UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

Coordenação de Pós-Graduação em Ciência da Computação

TESE DE DOUTORADO

AGRUPAMENTO DE FACES EM VÍDEOS DIGITAIS

EDUARDO SANTIAGO MOURA

ORIENTADORES

HERMAN MARTINS GOMES

JOÃO MARQUES DE CARVALHO

CAMPINA GRANDE, PARAÍBA

JULHO – 2016

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA

Coordenação de Pós-Graduação em Ciência da Computação

AGRUPAMENTO DE FACES EM VÍDEOS DIGITAIS

EDUARDO SANTIAGO MOURA

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários à obtenção do grau de Doutor em Ciência da Computação (DSc).

Área de concentração: Ciência da Computação

Linha de pesquisa: Metodologia e Técnicas da Computação

Herman Martins Gomes

João Marques de Carvalho

Orientadores

Campina Grande - Paraíba

Julho – 2016

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

M929a Moura, Eduardo Santiago.

Agrupamento de faces em vídeos digitais / Eduardo Santiago Moura. – Campina Grande, 2016.

129 f. : il. color.

Tese (Doutorado em Ciência da computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2016.

"Orientação: Prof. Dr. Herman Martins Gomes, Prof. Dr. João Marques de Carvalho".

Referências.

1. Computação – Vídeos Digitais. 2. Faces em Vídeos Digitais – Agrupamento. 3. Vídeos Digitais – Aglomerativo Hierárquico. 4. Avaliação de Agrupamento. I. Gomes, Herman Martins. II. Carvalho, João Marques de. III. Universidade Federal de Campina Grande – Campina Grande (PB). IV. Título.

CDU 004:621.397.4(043)

RESUMO

Faces humanas são algumas das entidades mais importantes frequentemente encontradas em vídeos. Devido ao substancial volume de produção e consumo de vídeos digitais na atualidade (tanto vídeos pessoais quanto provenientes das indústrias de comunicação e entretenimento), a extração automática de informações relevantes de tais vídeos se tornou um tema ativo de pesquisa. Parte dos esforços realizados nesta área tem se concentrado no uso do reconhecimento e agrupamento facial para auxiliar o processo de anotação automática de faces em vídeos. No entanto, algoritmos de agrupamento de faces atuais ainda não são robustos às variações de aparência de uma mesma face em situações de aquisição típicas. Neste contexto, o problema abordado nesta tese é o agrupamento de faces em vídeos digitais, com a proposição de nova abordagem com desempenho superior (em termos de qualidade do agrupamento e custo computacional) em relação ao estado-da-arte, utilizando bases de vídeos de referência da literatura. Com fundamentação em uma revisão bibliográfica sistemática e em avaliações experimentais, chegou-se à proposição da abordagem, a qual é constituída por módulos de pré-processamento, detecção de faces, rastreamento, extração de características, agrupamento, análise de similaridade temporal e reagrupamento espacial. A abordagem de agrupamento de faces proposta alcançou os objetivos planejados obtendo resultados superiores (no tocante a diferentes métricas) a métodos avaliados utilizando as bases de vídeos *YouTube Celebrities* (KIM et al., 2008) e *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Palavras-chave: agrupamento de faces em vídeos, agrupamento aglomerativo hierárquico, avaliação de agrupamento.

ABSTRACT

Human faces are some of the most important entities frequently encountered in videos. As a result of the currently high volumes of digital videos production and consumption both personal and profissional videos, automatic extraction of relevant information from those videos has become an active research topic. Many efforts in this area have focused on the use of face clustering and recognition in order to aid with the process of annotating faces in videos. However, current face clustering algorithms are not robust to variations of appearance that a same face may suffer due to typical changes in acquisition scenarios. Hence, this thesis proposes a novel approach to the problem of face clustering in digital videos which achieves superior performance (in terms of clustering quality and computational cost) in comparison to the state-of-the-art, using reference video databases according to the literature. After performing a systematic literature review and experimental evaluations, the current approach has been proposed, which has the following modules: preprocessing, face detection, tracking, feature extraction, clustering, temporal similarity analysis, and spatial re-clustering. The proposed approach for face clustering achieved the planned objectives obtaining better results (according to different metrics) than those presented by methods evaluated on the *YouTube Celebrities videos dataset* (KIM et al., 2008) and *SAIVT-Bnews videos dataset* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Keywords: video face clustering, hierarchical agglomerative clustering, clustering evaluation.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ARI	***A**djusted **R**and **I**ndex*
AUC	***A**rea **U**nder the **C**urve*
CP	***C**ovariance **P**rofile*
DCNN	***D**eep **C**onvolutional **N**eural **N**etwork*
F	***F**-Measure*
FDDB	***F**ace **D**etection **D**ata Set and **B**enchmark*
FM	***F**olks and **M**allows Index*
GT	***G**round-**T**ruth*
JI	***J**accard **I**ndex*
HAC	***H**ierarchical **A**gglomerative **C**lustering*
LDA	***L**inear **D**iscriminant **A**nalysis*
LFW	***L**abeled **F**aces in the **W**ild*
OpenCV	***Open** Source **C**omputer **V**ision (Intel Library)*
PCA	***P**rincipal **C**omponent **A**nalysis*
PDI	**P**rocessamento **D**igital de **I**magens
PE	**P**rotocolo de **E**studo
PosErr	***Pos**ition **Err**or*
PR	***P**recision-**R**ecall*
RI	***R**and **I**ndex*
RS	**R**evisão **S**istemática
VC	**V**isão **C**omputacional
SE	***S**tandard **E**rror*
SIFT	***S**cale-**I**nvariant **F**eature **T**ransform*
SURF	***S**peeded **U**p **R**obust **F**eatures*

Capítulo 1

Introdução

Neste capítulo, são introduzidos os elementos fundamentais utilizados na realização desta pesquisa. O foco é dado na apresentação e formalização do problema em estudo – agrupamento de faces em vídeo – bem como no contexto e nas principais motivações que justificam a importância deste estudo na área de Visão Computacional (VC).

Faces humanas são algumas das entidades mais importantes, frequentemente encontradas em vídeos, podendo ser consideradas como características semânticas de alto nível (ANTONOPOULOS, NIKOLAIDIS e PITAS, 2007). Devido à crescente produção de vídeos digitais, a extração automática de informações relevantes de vídeos pessoais tornou-se um tema de pesquisa ativo. Esforços existentes nesta área têm sido dedicados ao uso de reconhecimento facial para auxiliar o processo de anotação das pessoas. No entanto, algoritmos de agrupamento de faces atuais ainda não são robustos às grandes variações de aparência presentes em situações de aquisição reais (LIN et al., 2010).

Diante do exposto, nesta Tese de Doutorado são discutidas as pesquisas no âmbito do agrupamento de faces em vídeos digitais. Tais pesquisas serviram de inspiração para a composição da abordagem proposta, na qual objetivou-se à obtenção do melhor desempenho em relação ao estado-da-arte das técnicas destinadas ao agrupamento de faces, como aqueles presentes nos estudos de Bhatt et al. (2014), Mian (2013), Chen et al. (2012), Cui et al. (2012), dentre outros. Como consequência, é possível facilitar a extração de informações relevantes e o compartilhamento de coleções de vídeos de usuários, como ocorre nas bases de vídeos *YouTube Celebrity* (KIM et al., 2008) e *YouTube Faces* (WOLF et al., 2011).

As seções remanescentes deste capítulo estão divididas como segue. Na Seção 1.1, argumenta-se sobre as motivações para a pesquisa. A descrição do problema a ser resolvido é apresentada na Seção 1.2. Os objetivos a serem alcançados e a relevância desta pesquisa são apresentados na Seção 1.3. Finalmente, na Seção 1.4, é apresentada a estrutura deste documento.

1.1. Motivações

Vídeos digitais desempenham um papel cada vez mais importante no cotidiano dos indivíduos com aplicações que vão desde notícias, até entretenimento, pesquisa científica, segurança e vigilância. Devido ao fato de câmeras digitais, dispositivos móveis multimídia e mídias de armazenamento estarem cada vez mais acessíveis, a produção de vídeos em geral tem crescido significativamente. Portanto, há a necessidade do desenvolvimento de sistemas de recuperação de dados de vídeo mais eficientes e eficazes (TURAGA, VEERARAGHAVAN e CHELLAPPA, 2009). Como fator adicional para essa demanda, a quantidade de vídeos digitais tem proliferado rapidamente por meio de novas mídias, tais como a *Internet* e a *TV Digital* (CHOI, DE NEVE e RO, 2010). Como exemplo, o *website* de compartilhamento de vídeos mais popular no mundo, *YouTube*[1], no ano de 2009, armazenava cerca de 6.3 bilhões de vídeos, além de 20 horas de novos vídeos sendo postados a cada minuto (SCHROEDER, 2009; WEBSITE MONITORING BLOG, 2010). Nos últimos dois anos, o *website* obteve cerca de 800 milhões de visitantes por mês, com uma média de 48 horas de novos vídeos postados por minuto (LOS ANGELES TIMES, 2013; YOUTUBE ADVERTISE, 2014).

Um fator relevante em relação a vídeos é o fato destes fornecerem informações temporais, úteis ao rastreamento e ao reconhecimento de objetos (WECHSLER, 2006). Por outro lado, a grande quantidade de dados gerada por um vídeo pode tornar difícil a identificação de partes relevantes do mesmo, o que pode ser resolvido pela inclusão de alguma forma de estruturação. Essa estruturação pode ser feita pela divisão do fluxo de

[1] YouTube, disponível em: *http://www.youtube.com*

quadros de um vídeo em segmentos, considerando-se os seguintes aspectos: a identificação dos objetos presentes, a ordem temporal ou grupos de objetos semelhantes (OKAMOTO et al., 2002).

Um objeto em uma imagem é descrito quer por um conjunto de medidas quer pôr suas relações com outros objetos da imagem. Portanto, a organização em agrupamentos pode ser considerada como um dos modos mais fundamentais de compreensão e aprendizagem de objetos (JAIN, 1991). Um agrupamento constitui uma funcionalidade básica de análise de dados que fornece um resumo dos padrões de distribuição e correlação de dados em um conjunto, com aplicação em diversas áreas, tais como, processamento de imagens, compressão de dados, reconhecimento de padrões (JENSEN, LIN e OOI, 2007).

Dentre as aplicações de agrupamento de faces destacam-se: (i) a sumarização de vídeos de segurança (SONY et al., 2011); (ii) a identificação de indivíduos específicos em vídeos de noticiários (GAO, EKENEL e STIEFELHAGEN, 2011); (iii) a catalogação de cenas por atores presentes no vídeo (YAMAMOTO, YAMAGUCHI e AOKI, 2010); (iv) a indexação de vídeos baseada na atividade dos objetos (TURAGA, VEERARAGHAVAN e CHELLAPPA, 2009); (v) a listagem do elenco de atores (*cast list*) (ZHANG et al., 2009); (vi) o rastreamento e reconhecimento de faces em vídeos de baixa qualidade (KIM et al., 2008); (vii) o agrupamento de faces em vídeos digitais (TAO e TAN, 2008; HUANG, WANG e SHAO, 2008; ANTONOPOULOS, NIKOLAIDIS e PITAS, 2007); e (viii) a indexação e pesquisa de pessoas específicas (FOUCHER e GAGNON, 2007; RAMANAN, BAKER e KAKADE, 2007).

Diante do exposto, esta pesquisa objetivou inicialmente o estudo de abordagens para o agrupamento de objetos em vídeos digitais, a aplicação dessas abordagens ao contexto de faces, assim como a investigação de eventuais limitações e problemas em aberto presentes em sistemas de agrupamento de faces em vídeos digitais. Em seguida, a partir das técnicas e métodos levantados foi constituída a abordagem proposta. Por fim, realizou-se um estudo comparativo para validar a implementação das técnicas propostas, com base em métricas objetivas para avaliar a qualidade

dos resultados em bases de vídeos de referência da área.

1.2. Descrição do Problema

O reconhecimento de padrões tem como um de seus objetivos a classificação de objetos (padrões) em um número de categorias ou classes (THEODORIDIS e KOUTROUMBAS, 1999). No caso do reconhecimento de faces (quando se atribui um rótulo específico dentre um conjunto pré-definido), as imagens de faces são os objetos e as classes são seus nomes ou identificações. Dado um padrão, seu processo de reconhecimento/classificação do mesmo pode ser categorizado como: *supervisionado*, quando o padrão de entrada é identificado como um membro de uma classe pré-definida pelos padrões de treinamento, que são rotulados com suas classes; e *não supervisionado* ou *clustering*, quando o padrão é associado (agrupado) a uma classe que é aprendida com base na similaridade entre os padrões de treinamento, ou seja, o próprio sistema toma a decisão de criar novas classes ou agrupar classes pré-existentes.

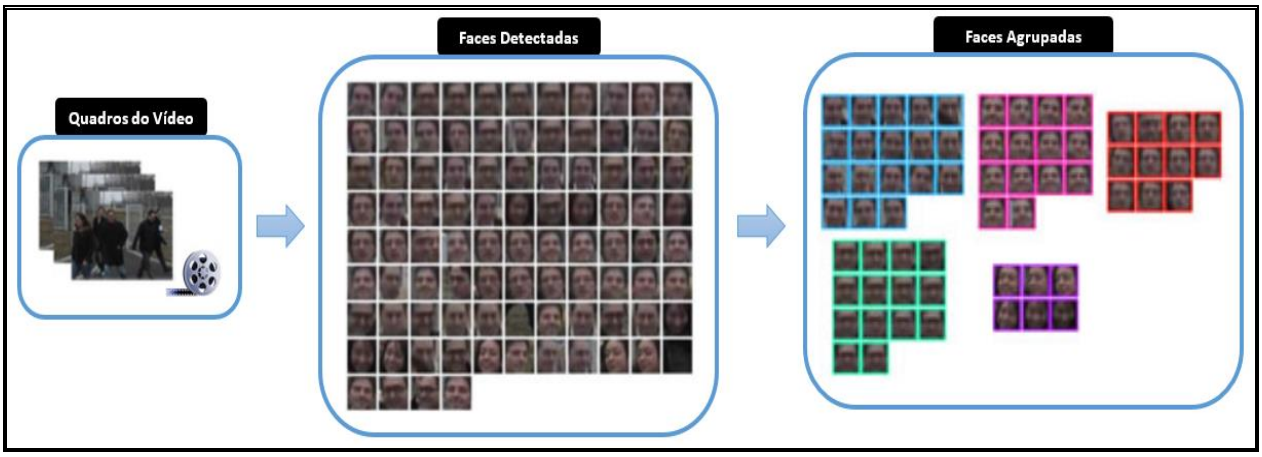
Dentre as aplicações de técnicas de reconhecimento supervisionado, destaca-se a identificação para controle de acesso, segurança e vigilância (ZHAO et al., 2003). Como aplicações de técnicas de reconhecimento não supervisionado, destacam-se o agrupamento de objetos, a mineração de dados e técnicas auxiliares a diagnósticos médicos (JAIN, 1991).

Nos métodos de agrupamento (*clustering*) de faces, as características faciais são utilizadas como identificadores que permitem agrupar várias imagens da face de uma mesma pessoa em um determinado grupo (*cluster*). A fim de extrair características de identificação, para uma dada pessoa, em uma determinada fotografia, primeiramente a região da face é detectada e, em seguida, são aplicados extratores de características na região detectada, tentando produzir uma representação única da face.

O problema do agrupamento de faces pode, portanto, ser formulado como segue. Dado um vídeo digital com quadros contendo faces humanas, deve-se agrupar as faces conforme sua similaridade entre as mesmas, sem um conhecimento prévio de qualquer uma das pessoas envolvidas. As faces

encontradas podem ter sido obtidas sob diferentes condições de imageamento, tais como pose, iluminação, expressão facial e oclusão parcial (vide Figura 1.1).

Figura 1.1 – Agrupamento de faces de vídeos.



Assim, o agrupamento de faces pode ser considerado como uma forma de classificação não supervisionada aplicada a um conjunto finito de objetos, cujo objetivo é agrupá-los em classes, de tal forma que objetos similares sejam colocados no mesmo grupo, enquanto objetos diferentes sejam colocados em grupos diferentes (ANTONOPOULOS, NIKOLAIDIS e PITAS, 2007).

1.3. Objetivos e Relevância

Torna-se evidente o contínuo interesse da indústria na área de agrupamento de faces, fato que pode ser observado pelo financiamento de pesquisas e investimentos em aplicativos de *software* comerciais por grandes empresas ao longo desta década, a exemplo do *Adobe Photoshop Elements*[2], *Google Picasa*[3], e *PittPatt FaceSort*[4]. Como exemplo local, pode-se citar o projeto *FaceRec*, ocorrido em 2013 no âmbito da cooperação técnico-científica firmada entre a *Hewlett-Packard* (HP) e a Universidade Federal de Campina Grande, com incentivos da Lei de Informática, o qual se insere no contexto do estudo de tais métodos.

Apesar de toda a pesquisa já desenvolvida, o agrupamento de faces humanas em vídeos, de forma automática, precisa e robusta, ainda constitui

[2] Adobe Photoshop Elements, disponível em: *http://www.adobe.com/br/products/photoshop-elements.html*
[3] Google Picasa, disponível em: *http://www.google.com/intl/pt-BR/picasa/*
[4] PittPatt FaceSort, disponível em: *http://pittpatt-facesort.software.informer.com/2.1/*

um problema em aberto. Dificuldades surgem devido a vários fatores, dentre os quais podem ser destacados: (i) diferenças de resolução da face na imagem; (ii) variações na escala e orientação da face; (iii) variações nas condições de iluminação; (iv) variações de pose; (v) variações de expressões faciais; e (vi) geração dos grupos de forma automática, sem a predeterminação da quantidade de grupos.

1.3.1. Objetivo Geral

O objetivo geral desta Tese de Doutorado é propor uma abordagem automática para o agrupamento de faces em vídeos digitais, fundamentada na coerência temporal entre quadros adjacentes, destinada à sumarização de trechos relevantes para os usuários e objetivando a obtenção de melhor desempenho, em comparação com estudos do estado-da-arte.

1.3.2. Objetivos Específicos

a) Investigar métodos e técnicas correntes destinadas ao agrupamento de faces em vídeos digitais;

b) Investigar e extrair características locais e globais que auxiliem a resolução dos problemas anteriormente mencionados;

c) Propor algoritmos mais eficientes para o agrupamento de faces (em termos de qualidade do agrupamento e custo computacional) em relação àqueles existentes na literatura da área;

d) Elaborar uma abordagem de agrupamento de faces por meio de um processo de aprendizagem não-supervisionada, a ser utilizado em tarefas que envolvam um número indeterminado de grupos;

e) Implementar os diferentes módulos da abordagem proposta visando à realização de avaliações experimentais e a publicação de resultados intermediários de cada módulo e da abordagem integrada;

f) Validar os resultados obtidos a partir de experimentos com métricas de avaliação objetivas e comparar a abordagem proposta com soluções existentes.

1.4. Estrutura da Tese

O presente documento é composto por um total de cinco capítulos, incluindo-se o presente capítulo. No **Capítulo 2**, apresenta-se um panorama das técnicas de agrupamento de faces em vídeos, a partir de uma revisão sistemática de pesquisas relevantes da área.

Descrevem-se, também, técnicas empregadas na tarefa específica de extração de características para a identificação pessoal, métricas de similaridade, estratégias comumente empregadas para resolver o problema objeto de estudo, métricas de avaliação da qualidade de agrupamento, bem como se delimita o escopo do presente estudo.

No **Capítulo 3,** detalha-se a abordagem proposta para a solução do problema de agrupamento de faces em vídeos, incluindo-se a arquitetura geral, o fluxo de processamento e o funcionamento de cada módulo da técnica proposta.

No **Capítulo 4,** reúnem-se os experimentos realizados e uma discussão dos resultados obtidos. A apresentação dos experimentos contempla os testes realizados após a proposição da abordagem descrita no Capítulo 3.

Finalmente, no **Capítulo 5**, apresenta-se uma breve discussão do que foi exposto na Tese de Doutorado, as conclusões finais e as propostas para pesquisas futuras.

Capítulo 2

Trabalhos Relacionados

Neste capítulo, apresenta-se um levantamento e análise de estudos nos quais são investigados ou propostos métodos que auxiliaram no amadurecimento de uma nova solução para o problema em estudo – agrupamento de faces em vídeos. Este capítulo está organizado como segue. Inicialmente, a metodologia para a pesquisa bibliográfica realizada é apresentada. Em seguida, tem-se a revisão propriamente dita, em que os estudos foram organizados em duas categorias: abordagens com base em conjuntos de quadros e abordagens com base em sequências de quadros.

2.1. Metodologia de Pesquisa Bibliográfica

A sistemática de pesquisa bibliográfica empregada na execução desta pesquisa objetivou determinar o estado-da-arte sobre o problema em questão, de forma a identificar o que foi publicado sobre o assunto nos últimos 5 anos, que aspectos já foram abordados e quais as lacunas existentes na literatura, de maneira a possibilitar a delimitação do problema a ser estudado, além de prover uma estruturação conceitual que dará sustentação ao desenvolvimento da pesquisa.

Uma pesquisa bibliográfica pode ser vista como um estudo experimental e, como tal, é caracterizado por um teste de hipótese sobre formas de análise quantitativa, qualitativa e semi-quantitativa. Tais estudos experimentais podem ser classificados em (KITCHENHAM e CHARTERS, 2007):

- Estudos primários: têm como objetivo investigar uma questão de pesquisa específica, por exemplo, estudos de caso, experimentos ou pesquisas de opinião;

- Estudos secundários: revisam os estudos primários relativos a uma ou mais questões de pesquisa, com o objetivo específico de sintetizar as evidências relacionadas a tais questões de pesquisa, tais como, revisões e mapeamentos sistemáticos;
- Estudos terciários: correspondem a revisões sistemáticas de revisões sistemáticas, com o objetivo de responder a uma questão de pesquisa mais abrangente, buscando-se elevar seu grau de qualidade.

As revisões bibliográficas são geralmente realizadas sob a forma de estudos primários, com pouca ou nenhuma sistematização, tornando-as passíveis de produzir resultados enviesados, duplicados e com pouco valor para a comunidade científica (MAFRA e TRAVASSOS, 2006). Resultados fundamentados neste cenário podem tornar oneroso o processo de revisão, tanto para os pesquisadores, devido ao tempo e esforço empregado, quanto para as instituições, que fornecem os recursos para a pesquisa.

Revisão Sistemática (RS) é uma metodologia de estudo secundária que objetiva realizar um levantamento formal e consistente do estado-da-arte, a partir de planejamento e execução criteriosos, permitindo avaliar e interpretar as publicações científicas relevantes para uma questão particular de pesquisa, um tópico de área ou um fenômeno de interesse (KHAN et al., 2001; PAI et al., 2004). Revisões sistemáticas são concebidas para serem metódicas, explícitas e passíveis de reprodução, com a potencialidade de elevar a qualidade dos levantamentos bibliográficos de pesquisas científicas, dado que se trata de uma metodologia utilizada para delimitar a abrangência de um estudo e garantir o acesso aos dados de interesse a todos os pesquisadores interessados no problema objeto de estudo (SAMPAIO e MANCINI, 2007).

Desta forma, o processo de pesquisa bibliográfica realizado nesta tese foi conduzido segundo uma sequência metodologicamente bem definida de etapas, de acordo com um protocolo de estudo previamente planejado, a fim de organizar ideias, conceitos, métodos, técnicas que dizem respeito ao problema objeto de estudo. Isto possibilitou o mapeamento de limitações

em estudos já realizados e de problemas em aberto, assim como, potenciais soluções existentes no estado-da-arte.

Buscou-se realizar um levantamento e estruturação do estado da arte sobre Agrupamento de Faces em Vídeos a partir da concepção de um Protocolo de Estudo (PE), baseado em uma revisão bibliográfica da literatura da área, a fim de organizar ideias, conceitos, métodos, técnicas que dizem respeito ao problema objeto de estudo. No Apêndice B, apresenta-se o PE composto das questões de pesquisa, a metodologia empregada no processo de busca, os critérios de inclusão, exclusão e avaliação de qualidade dos trabalhos relacionados.

Na Seção 2.2, descrevemos de forma detalhada o estado-da-arte sobre Agrupamento de Faces em Vídeos. Por fim, na Seção 2.3, são apresentadas as conclusões obtidas a partir da execução da pesquisa bibliográfica.

2.2. Revisão sobre Agrupamento de Faces em Vídeos

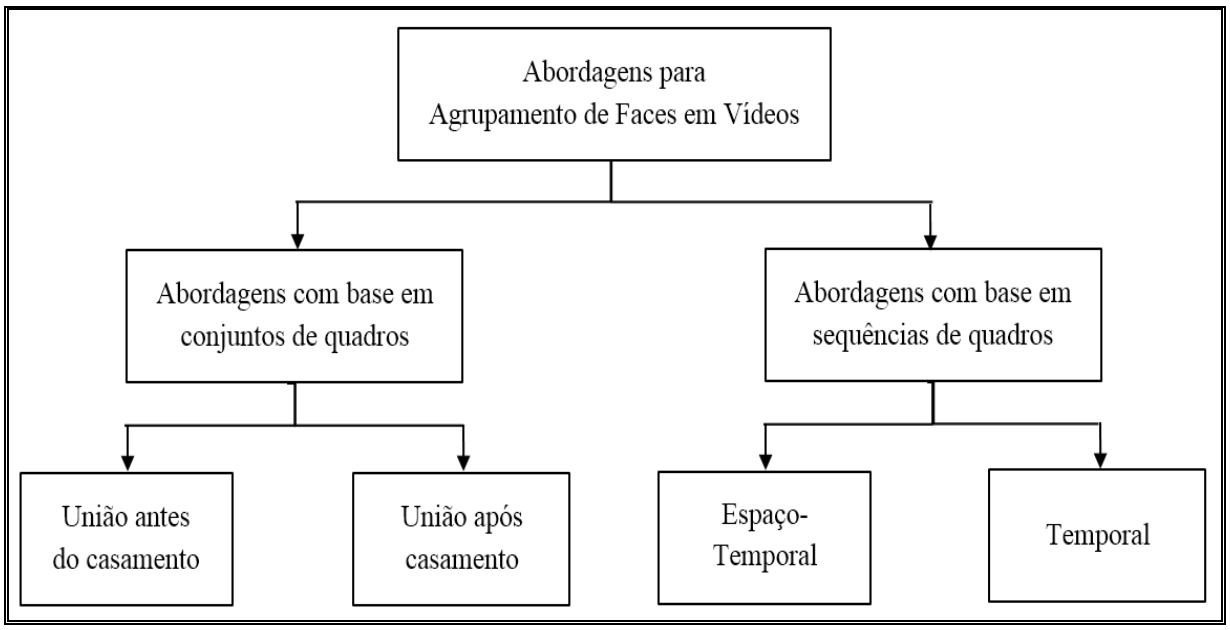
Conforme exposto no Capítulo 1, constata-se uma ênfase dada nos últimos anos ao agrupamento de objetos, em especial faces humanas, em vídeos digitais. Vídeos fornecem abundante informação que pode ser aproveitada para tratar problemas ocasionados por variações de pose, iluminação e expressão facial, bem como melhorar o desempenho do reconhecimento facial.

Adicionalmente, o reconhecimento de indivíduos em vídeos possui vantagens e desvantagens em comparação com imagens estáticas. Contudo, dado que a aquisição de vídeos não é sempre realizada em ambiente controlado, a presença de ruídos, tais como, pose, iluminação, oclusão é significativamente maior, por outro lado, a quantidade de informação disponível em um vídeo é maior do que a informação disponível ao se comparar duas imagens estáticas. Vídeos fornecem dados adicionais em termos de múltiplos quadros e informação temporal se comparados a imagens estáticas. Tais informações podem ser utilizadas para melhorar o

desempenho dos sistemas de reconhecimento facial e podem prover robustez a grandes variações de pose, expressão facial e iluminação.

As abordagens publicadas sobre o tema de agrupamento de faces em vídeos podem ser categorizadas em dois grupos: abordagens com base em conjuntos de quadros e abordagens com base em sequência de quadros, dependendo de quais propriedades do vídeo são utilizadas, conforme ilustrado na Figura 2.1.

Figura 2.1 – Taxonomia de agrupamento de faces em vídeos.



Abordagens com base em conjuntos de quadros (e.g., OTTO, WANG e JAIN, 2016; SCHROFF, KALENICHENKO e PHILBIN, 2015; CAO et al., 2015A; ANOOP et al, 2012; CUI et al., 2012; WOLF et al., 2011; HU et al., 2011; HARANDI et al., 2011; WANG et al., 2008; ANTONOPOULOS, NIKOLAIDIS e PITAS, 2007; NISHIYAMA et al., 2007; FOUCHER e GAGNON, 2007; FUKUI e YAMAGUCHI, 2007) tratam os vídeos como coleções desordenadas de imagens e aproveitam a multiplicidade de observações, enquanto que as abordagens com base em sequências de quadros explicitamente utilizam a informação temporal para aumentar a eficiência ou permitir o reconhecimento em condições adversas.

Embora abordagens com base em conjunto de quadros não dependam da ordem temporal das faces presentes, estas exploram a quantidade e variedade de observações para atingir robustez em condições de visualização degradadas. Tais abordagens diferem em termos de como é

realizada a união de informações sobre as faces presentes nos quadros antes ou após o casamento (*matching*) de faces individualmente, para posterior agrupamento. Por exemplo, todo o conjunto de observações pode ser modelado como um centroide ou uma distribuição de probabilidade, objetivando aumentar a robustez a variações de pose, iluminação e de expressão facial. Alternativamente, a união de informações pode ocorrer em função do nível e do *ranking* de similaridade do casamento (*matching*) de subconjuntos de quadros, de maneira que tais subconjuntos possam ser representativos por conter potenciais observações robustas a variações de pose, iluminação e expressão facial.

Em contraste com as abordagens com base em conjuntos de quadros, as abordagens com base em sequências de quadros (ANANTHARAJAH et al., 2015; ZHOU et al., 2015; TANG et al.,2015; CAO et al., 2015B; TAPASWI et al., 2014; BHATT et al., 2014; MIAN, 2013; CHEN et al., 2012; SONY et al., 2011; GAO, EKENEL e STIEFELHAGEN, 2011; YAMAMOTO, YAMAGUCHI e AOKI, 2010; TURAGA, VEERARAGHAVAN e CHELLAPPA, 2009; ZHANG et al., 2009; KIM et al., 2008; TAO e TAN, 2008; HUANG, WANG e SHAO, 2008; RAMANAN, BAKER e KAKADE, 2007) utilizam explicitamente informações espaço-temporais em função da aparência e movimento para auxiliar o casamento entre faces semelhantes. Métodos baseados em sequências de quadros podem, também, auxiliar na tarefa de rastreamento de faces melhorando o desempenho do reconhecimento em condições degradadas, principalmente, no cenário de oclusão parcial ou total.

Na Seção 2.2.1 são detalhados os estudos categorizados como abordagens com base em conjunto de quadros. De maneira semelhante, na Seção 2.2.2 são discutidos os estudos categorizados como abordagens com base em sequência de quadros. Em ambas as seções, o critério de apresentação adotado para a apresentação das publicações foi a ordem cronológica reversa, ou seja, do artigo mais atual para o mais antigo, de modo a detalhar e enfatizar os trabalhos mais recentes, uma vez que, apresentaram resultados superiores. Por fim, na Seção 2.3, são apresentadas as considerações e as conclusões resultantes da análise dos trabalhos relacionados.

2.2.1. Abordagens com base em Conjuntos de Quadros

As abordagens com base em conjuntos de quadros tratam o problema de agrupamento de faces em vídeos em termos do casamento de conjuntos de múltiplas amostras. Tais técnicas unem a informação sobre o conjunto de amostras antes ou após o casamento individual de faces. A união da informação permite que algoritmos de agrupamentos de faces com base em conjuntos de quadros obtenham maior acurácia de reconhecimento, maior robustez à presença de ruídos e maior eficiência em relação aos algoritmos de agrupamento de faces com base em imagens estáticas. Alguns dos métodos revisados nesta seção foram avaliados utilizando bases de imagens estáticas, o que é justificado pelo fato dos mesmos não utilizarem informação temporal, processando cada quadro independentemente dos demais.

Assim, as abordagens com base em conjuntos de quadros podem ser divididas em duas subcategorias:

- **União antes do casamento (*matching*):** as características extraídas de cada imagem de face são agregadas antes do casamento. Os valores de cada pixel de uma imagem são utilizados para a formação de um vetor de características e a concatenação de vetores de diferentes quadros produz um único vetor com a informação de um conjunto inteiro. Uma desvantagem desta forma de representação deriva da sua sensibilidade ao número de faces e à ordem com que os vetores são concatenados. Em contraste, nos métodos de superresolução a meta é recuperar o conteúdo de alta frequência da imagem a partir dos quadros agregados, com o objetivo de construir imagens em alta resolução. Algumas técnicas de modelagem 3D também objetivam extrair dados de múltiplos quadros, apenas com o intento de aproximar a estrutura geométrica da face para obter invariância a pose. Adicionalmente, todo o conjunto de faces pode ser representado por subespaços lineares ou centroides não-lineares construídos com métricas bem definidas

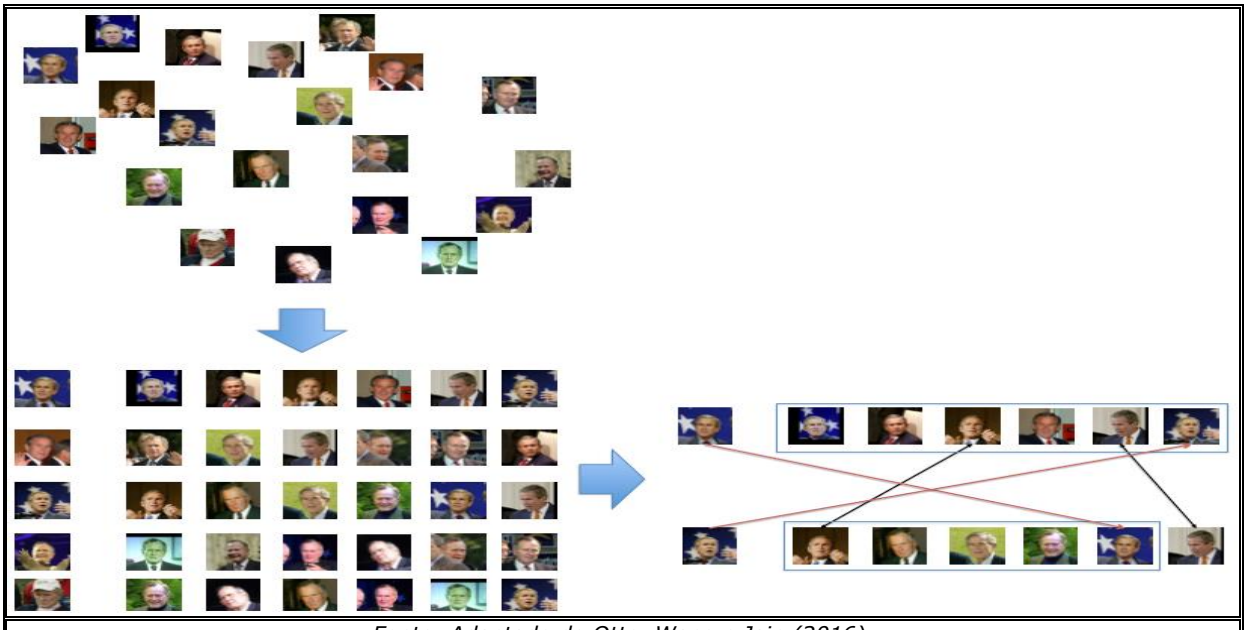
que mensuram as distâncias entre os conjuntos ou as variações que eles possuem em comum. Os artigos enquadrados nesta categoria foram dos seguintes autores: Schroff, Kalenichenko e Philbin (2015), Anoop et al. (2012), Cui et al. (2012), Wolf et al. (2011), Hu et al. (2011), Wang et al. (2008) e Fukui e Yamaguchi (2007);

- **União após o casamento (*matching*):** variações de pose, iluminação e expressões faciais dificultam o reconhecimento facial, por afetar a aparência da face. Conjuntos de imagens podem ser amostrados por meio de algoritmos de seleção de quadros para aumentar a probabilidade de que os conjuntos de treinamento e teste tenham composições semelhantes em relação aos fatores de perturbação. Algumas técnicas utilizam modelos de face 3D para obter invariância à pose, sintetizando imagens de treinamento com orientações similares àquelas das faces pertencentes ao conjunto de teste. Esses métodos podem ser complementados por técnicas de união por pontuação (*score*) ou *ranking* que integram informações sobre as imagens de treinamento e teste para produzir uma única decisão sobre o casamento. Na união por pontuação, o casamento por meio dos quadros de teste é via soma, multiplicação ou considerando-se a menor ou a maior pontuação para cada conjunto de treinamento. A identidade estimada corresponde ao conjunto de treinamento de maior pontuação combinada. Na união por *ranking*, os conjuntos de treinamento são classificados por sua pontuação em ordem decrescente para cada quadro. O conjunto de treinamento com a menor soma de rankings sobre o conjunto de quadros é considerado como a identidade estimada. Os artigos enquadrados nesta categoria foram dos seguintes autores: Otto, Wang e Jain (2016), Cao et al. (2015A), Harandi et al. (2011), Antonopoulos, Nikolaidis e Pitas (2007), Foucher e Gagnon (2007) e Nishiyama et al. (2007).

Otto, Wang e Jain (2016) abordaram o problema do agrupamento de

faces em conjuntos de quadros. Em cenários de larga escala, o número de faces da coleção pode chegar a uma ordem de centenas de milhões de imagens, enquanto o número de grupos pode variar entre poucas centenas de milhões, o que implica dificuldades em termos da complexidade, do tempo de execução e da qualidade do agrupamento. Um novo algoritmo de agrupamento, denominado *Rank-Order*, foi proposto e os autores declaram terem obtido a escalabilidade desejada e melhor acurácia de agrupamento em comparação com algoritmos conhecidos, como *K-Means* e *Spectral Clustering*. O processo de extração de características consiste na detecção de 68 pontos fiduciais, seguida de um método de árvores de regressão proposto por Kazemi e Sullivan (2014). Por fim, a imagem é normalizada com base nos *keypoints* detectados e rotacionada no plano, em função do ângulo entre os olhos e o ponto central da imagem. O algoritmo de agrupamento proposto é uma variação do agrupamento aglomerativo hierárquico, o qual utiliza uma lista de vizinhos mais próximos como medida de distância que é atualizada iterativamente a cada fusão entre grupos mais próximos, conforme ilustrado na Figura 2.2. O método apresentou o valor de 87% para a métrica de avaliação de agrupamento *F-Measure* na base de imagens LFW (HUANG et al., 2007) e valores de 71%, 79% e 67% para as métricas *F-Measure*, *Precision* e *Recall*, respectivamente, na base de vídeos *YouTube Faces* (WOLF et al., 2011).

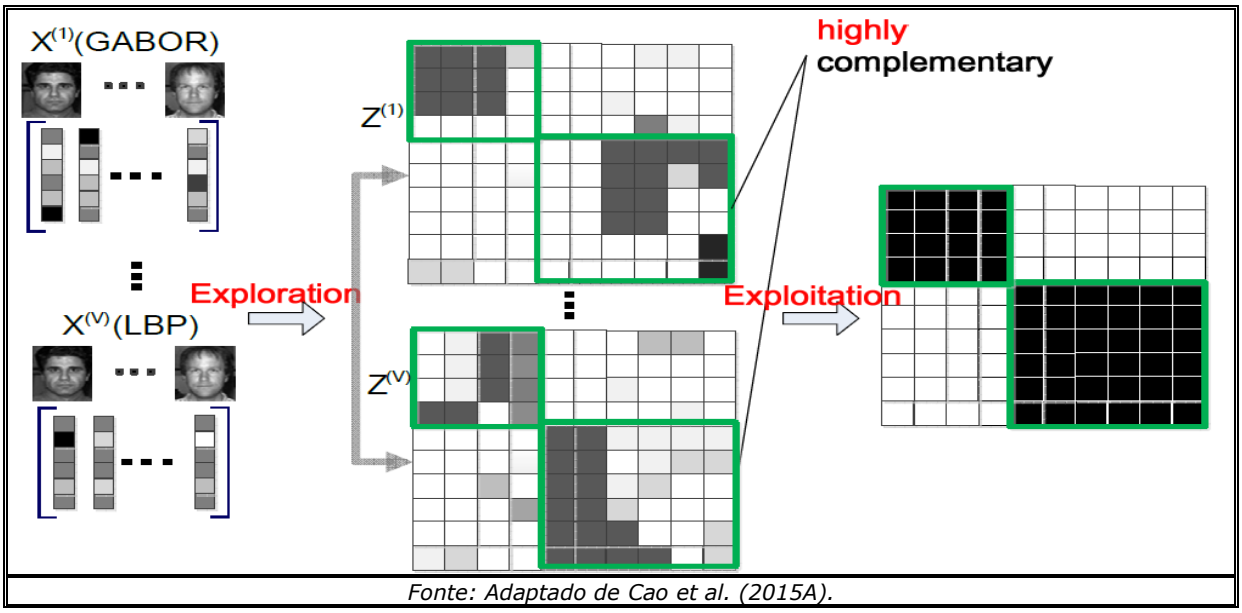
Figura 2.2 – Ilustração da abordagem proposta por Otto, Wang e Jain (2016).



Fonte: Adaptado de Otto, Wang e Jain (2016).

Cao et al. (2015A) focam seu estudo em como melhorar o agrupamento *multi-view,* explorando informações complementares entre características *multi-view*. Os autores propõem uma nova abordagem, denominada *Diversity-induced Multi-view Subspace Clustering* (DiMSC), que estende o subespaço existente do agrupamento em um domínio *multi-view* e utiliza o critério *Hilbert Schmidt Independence Criterion* (HSIC) como termo de diversidade para explorar a complementaridade de representações *multi-view*. O propósito é reduzir a redundância das representações e melhorar a qualidade do agrupamento, conforme ilustrado na Figura 2.3. Experimentos foram realizados em três bases públicas de conjuntos de quadros Yale (BELLHUMER, HESPANHA e KRIEGMAN, 1996), Yale B (GEORGHIADES, BELHUMEUR e KRIEGMAN, 2001) e ORL (SAMARIA e HARTER, 1994) com valores da métrica de avaliação de agrupamento *Adjusted Rand Index* (ARI) de 53,50%, 45,30% e 80,20%, respectivamente.

Figura 2.3 – Fluxograma da abordagem proposta por Cao et al. (2015A).



Fonte: Adaptado de Cao et al. (2015A).

Schroff, Kalenichenko e Philbin (2015) propuseram o método *FaceNet*, que diretamente aprende um mapeamento de imagens de faces para um compacto espaço Euclidiano no qual as distâncias correspondem diretamente a uma medida de similaridade facial. O método utiliza rede convolucional profunda (*deep convolutional neural network*). Para treinar a rede, é utilizada a correspondência de padrões faciais *matching* / *non-matching* com base no método *Large Margin Nearest Neighbor* (LMNN),

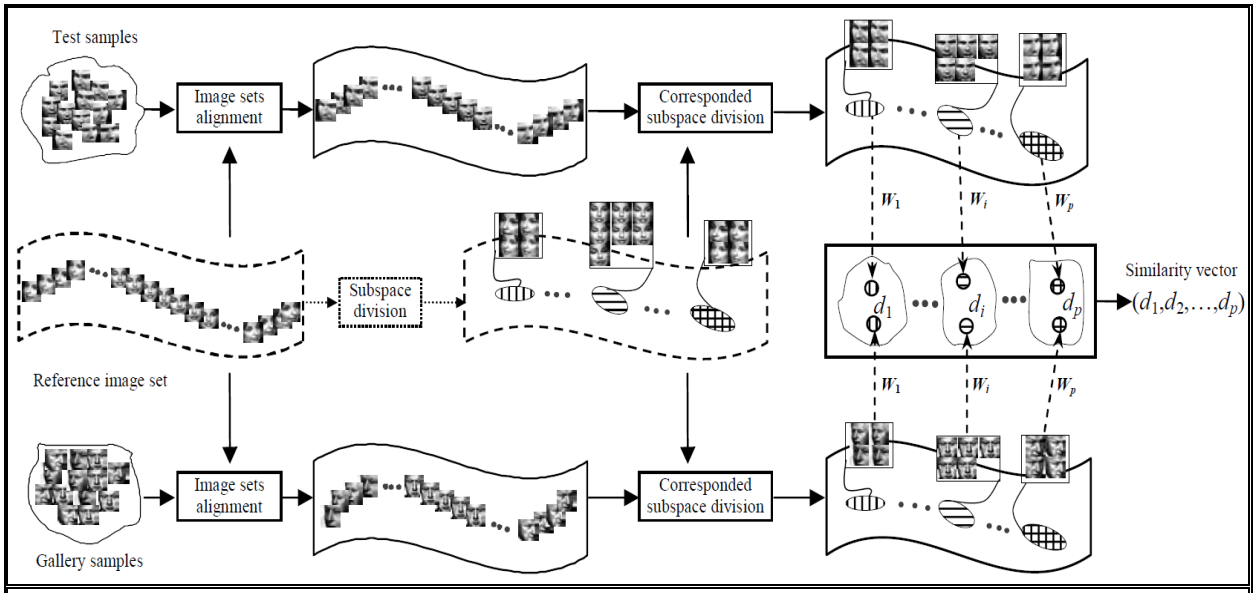
proposto por Weinberger, Blitzer e Saul (2006). A abordagem apresentou o valor de acurácia de 99,63% na base de imagens LFW (HUANG et al., 2007) e valor de acurácia de 95,12% na base de vídeos *YouTube Faces* (WOLF et al., 2011).

No estudo de Anoop et al. (2012) foi tratado o problema de se extrair uma assinatura representativa de entidades semelhantes (faces de indivíduos) por meio de descritores de covariância. De acordo com os autores, descritores de covariância podem eficientemente representar objetos e são robustos às variações de pose e escala, uma vez que compartilham uma estrutura geométrica comum que pode ser extraída por uma diagonalização conjunta. Os autores propuseram o *Covariance Profile* (CP), um descritor de assinatura que representa compactamente um conjunto de objetos similares. A ideia por trás do CP é que as mesmas direções principais são compartilhadas por objetos semelhantes. Tais direções são obtidas simultaneamente pela diagonalização das matrizes de covariância correspondendo aos objetos individuais. A representação facial é composta de características Gabor (RIESENHUBER e POGGIO, 1999), as quais constituem a base do CP. O agrupamento é realizado com base no *Spectral clustering* atuando sobre as matrizes de covariância. Experimentos foram realizados nas bases *Sitcom* (WOLF et al., 2011) e *YouTube Celebrities* (KIM et al., 2008), nos quais foram obtidos os valores 90,76% e 80,35% para a métrica *Average Purity* (P_w), respectivamente.

O método proposto por Cui et al. (2012) objetiva alinhar dois conjuntos de quadros contendo faces a um conjunto de referência bem definido e pré-estruturado para um número de modelos faciais locais antes do casamento. Isto é, dado dois conjuntos de imagens de faces, enquanto ambos estejam alinhados com o conjunto de referência, os mesmos estão mutuamente alinhados e bem estruturados. Assim, a similaridade entre eles pode ser calculada apenas comparando-se os modelos faciais locais correspondentes, ao invés de considerar todos os pares. Para se alinhar um conjunto de quadros com o conjunto de referência, são consideradas três restrições: (a) termo de correspondência explorando ângulos principais; (b) consistência geométrica estruturante utilizando reconstrução afim de pesos

invariantes; e (c) preservação local de relação de vizinhança, conforme ilustrado na Figura 2.4. Experimentos foram realizados em três bases públicas de conjuntos de quadros Honda/UCSD (LEE et al., 2005), CMU MoBo (GROSS e SHI, 2001) e *YouTube Celebrities* (KIM et al., 2008) com valores da métrica *Average Purity* (P_w) de 98,90%, 95,00% e 74,60%, respectivamente.

Figura 2.4 – Fluxograma da abordagem proposta por Cui et al. (2012).

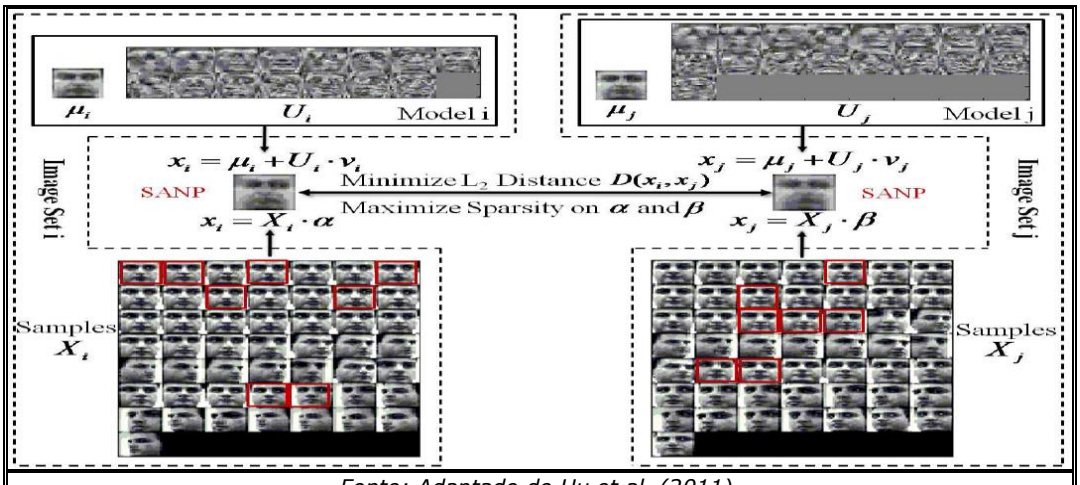


Fonte: Adaptado de Cui et al. (2012).

Wolf et al. (2011) projetaram uma abordagem com base em similaridade entre conjuntos para a comparação de quadros entre dois vídeos contendo imagens de faces, objetivando determinar se as faces presentes nos dois conjuntos de quadros são de uma mesma pessoa, ignorando semelhanças devido à pose, às condições de iluminação e à visualização. De maneira a acentuar as semelhanças de identidade, um classificador foi treinado para os indivíduos pessoas de cada sequência de vídeo, em que um conjunto de quadros é modelado por uma combinação de tais classificadores. O cerne da abordagem considera que a similaridade é assimétrica e utiliza a comparação entre os classificadores de um conjunto para determinar se o outro conjunto é o mais parecido com o atual. Foi realizado um experimento na base de imagens *YouTube Faces* (WOLF et al., 2011) com uma acurácia de 72,60%. A taxa obtida, não tão elevada, suscita complexidade da base de dados, a qual é composta por 3.425 vídeos, selecionados do *YouTube,* com 1.595 diferentes pessoas.

Hu et al. (2011) propuseram um algoritmo para a classificação de conjuntos de quadros, sendo empregada a tupla como representação: [um número de amostras de imagem, a sua média e um modelo *affine hull*]. Tal modelo é utilizado para contabilizar as aparências faciais não vistas sob a forma de combinações afim das amostras de imagem. Os autores introduziram a distância inter-classe (*between-class*) denominada *Sparse Approximated Nearest Point* (SANP) que mensura a dissimilaridade entre dois conjuntos de quadros pela distância entre seus pontos mais próximos, que pode ser esparsamente aproximada a partir das amostras faciais do respectivo conjunto de imagens, conforme ilustrado na Figura 2.5. Essa abordagem reforça a separabilidade dos coeficientes das amostras em vez dos coeficientes do modelo e otimiza conjuntamente os pontos mais próximos, bem como suas aproximações esparsas. Experimentos foram realizados em três bases públicas de conjuntos de quadros *Honda/UCSD* (LEE et al., 2005), *CMU MoBo* (GROSS e SHI, 2001) e *YouTube Celebrities* (KIM et al., 2008) com valores da métrica *Average Purity* (P_w) de 92,31%, 97,08% e 65,03%, respectivamente.

Figura 2.5 – *Sparse Approximated Nearest Point* (SANP) de dois conjuntos de quadros de imagens.

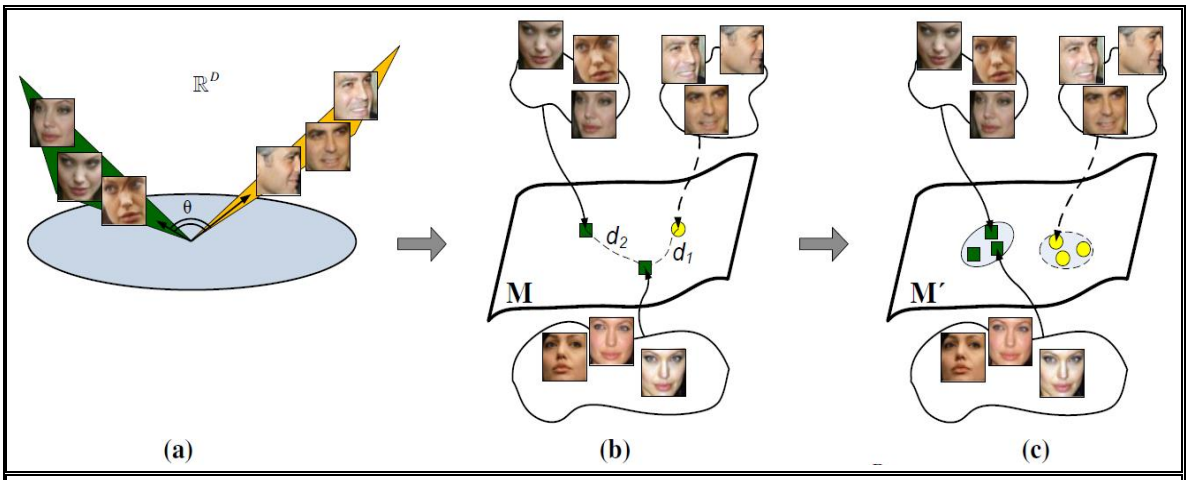


Fonte: Adaptado de Hu et al. (2011).

Harandi et al. (2011) argumentaram em seu trabalho que uma maneira conveniente de lidar com conjuntos de quadros de imagens de faces é representá-los como pontos em superfícies ou espaços topológicos denominados de *Grassmannian manifolds*. Os autores propuseram uma abordagem de análise discriminativa com base em grafos introduzindo as similaridades intra-classe (*within-class*) e inter-classe (*between-class*) para

caracterizar a compacidade intra-classe e a separabilidade inter-classe de forma que a estrutura geométrica dos dados possa ser explorada, conforme ilustrado na Figura 2.6. Experimentos realizados em vários conjuntos de bases de imagens (*CMU PIE* (SIM, BAKER e BSAT, 2003), *BANCA* (BAILLY-BAILLIÉRE et al., 2003) e *CMU Mobo* (GROSS e SHI, 2001) com 75,80%, 68.73% e 89,92% de acurácia, respectivamente) demonstraram que o algoritmo proposto obtém melhores resultados na acurácia de discriminação, em comparação com três métodos recentes: *Grassmann Discriminant Analysis* – GDA (HAMM e LEE, 2008), *Kernel* GDA (WANG e SHI, 2009) e a versão *kernel* do algoritmo *Affine Hull Image Set Distance* – *Kernel* AHISD (CEVIKALP e TRIGGS, 2010).

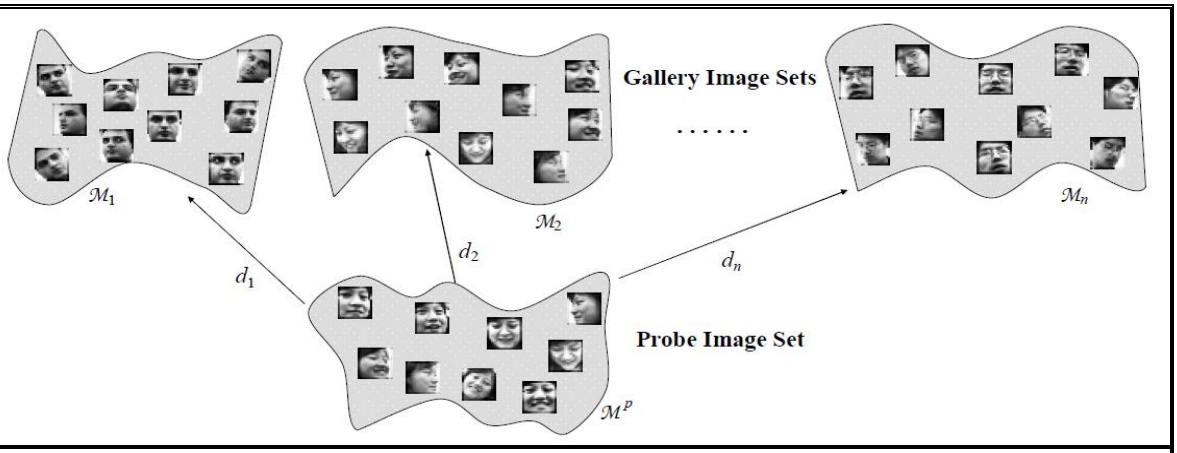
Figura 2.6 – Ilustração conceitual da abordagem proposta por Harandi et al. (2011). (a) Conjuntos de imagens podem ser descritos por seus subespaços lineares. Para comparar dois subespaços lineares, os ângulos principais entre eles podem ser utilizados. (b) Subespaços lineares podem ser representados como pontos no Grassmannian manifold M. (c) Determinação da similaridade entre os manifold.



Fonte: Adaptado de Harandi et al. (2011).

Wang et al. (2008) abordaram o problema da classificação de conjuntos de imagens extraídas de vídeos, segundo o qual cada conjunto contém imagens de faces pertencentes à mesma classe, mas cobrindo grandes variações de pose e iluminação. Tal classificação é realizada por meio do cálculo de uma nova medida de distância entre centroides (*Manifold-Manifold Distance* – MMD) não lineares que são originados por um conjunto de imagens. Os centroides são expressos por uma coleção de modelos locais lineares descritos por um subespaço, conforme ilustrado na Figura 2.7. Experimentos foram realizados em duas bases públicas de imagens *Honda/UCSD* (LEE et al., 2005) e *CMU MoBo* (GROSS e SHI, 2001) com taxas de reconhecimento de 96,9% e 93,6%, respectivamente.

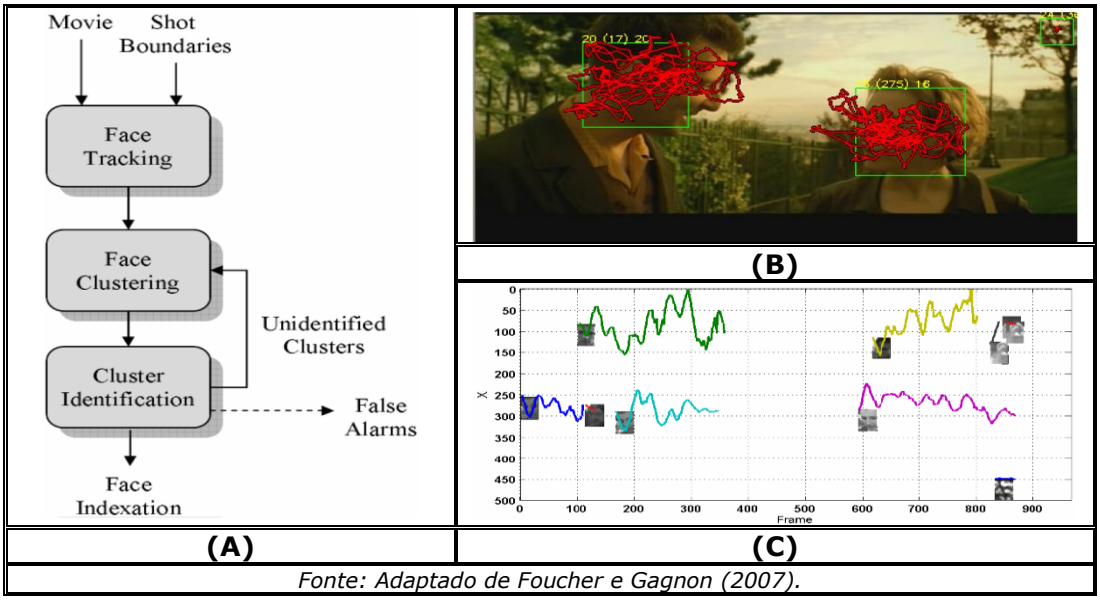
Figura 2.7 – Exemplo do cálculo da distância entre centroides do método de Wang et al. (2008).



Fonte: Adaptado de Wang et al. (2008).

Foucher e Gagnon (2007) elaboraram um sistema de indexação de vídeos de longa duração, que leva em consideração a presença de faces semelhantes. O sistema contém um módulo de detecção de faces frontais e semi-frontais, implementado como uma cascata de classificadores fracos (VIOLA e JONES, 2001). As faces detectadas são rastreadas por meio de um filtro de partículas e pelas trajetórias geradas pelo mesmo e, em seguida, o agrupamento de faces é realizado, conforme fluxograma ilustrado na Figura 2.8. Resultados experimentais permitiram verificar que o sistema proposto apresentou uma acurácia de apenas 25% quando aplicado a um vídeo com um grande número de faces diferentes presentes. Uma limitação do sistema proposto é que este é incapaz de detectar faces com grandes variações de pose (e.g., perfil lateral ou *full-profile*).

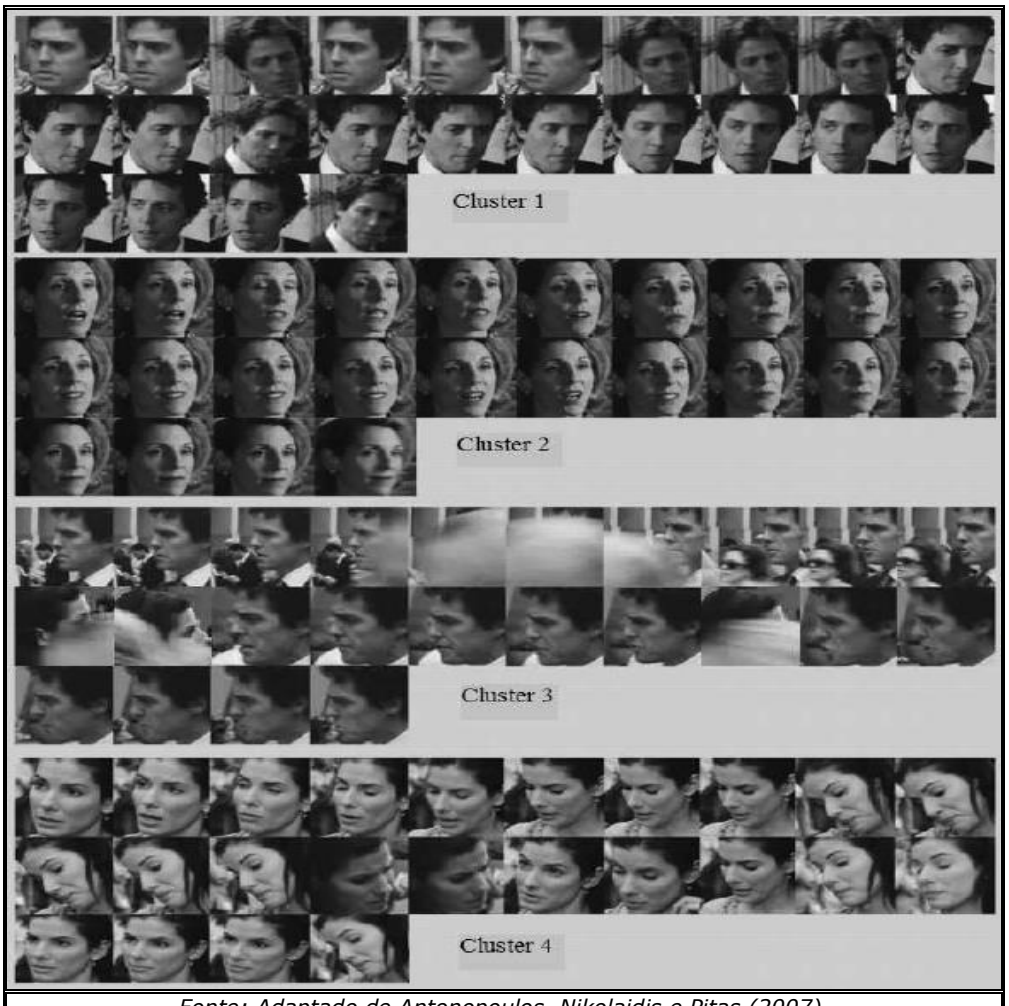
Figura 2.8 – (A) Arquitetura proposta por Foucher e Gagnon (2007). (B) Rastreamento de dois atores em movimento (em vermelho face tracks e regiões em verde resultado do filtro de partículas). (C) Face tracks por quadro e posição espacial.



Fonte: Adaptado de Foucher e Gagnon (2007).

Antonopoulos, Nikolaidis e Pitas (2007) desenvolveram um método para agrupar faces em quadros de vídeos que utiliza uma matriz de similaridade usada como entrada para um algoritmo de agrupamento hierárquico. Três métricas de avaliação bem conhecidas foram empregadas para avaliar a qualidade do agrupamento resultante, a saber: (i) *F-Measure*; (ii) Entropia Total (OE); e (iii) Estatística *Γ*. Resultados experimentais no vídeo *Two Weeks Notice*, apresentaram taxas de 0,8763 para a *F-Measure*, 0,7977 para a medida OE e 0,094 para a estatística Γ. Os autores relatam que tais resultados podem ser considerados robustos em relação a variações de escala, pose e iluminação, conforme ilustrado na Figura 2.9.

Figura 2.9 – Resultados visuais do método proposto por Antonopoulos, Nikolaidis e Pitas (2007).

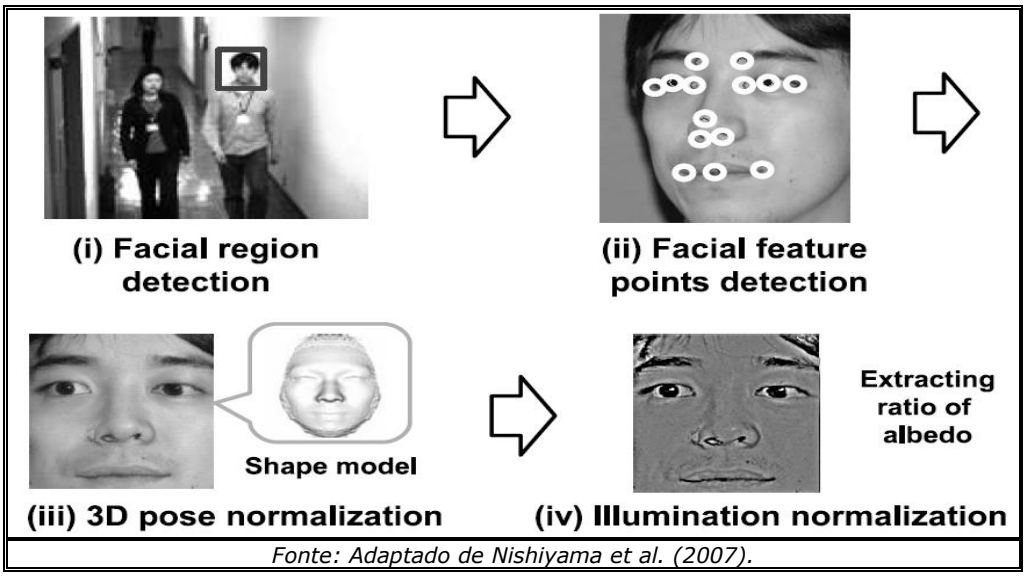


Fonte: Adaptado de Antonopoulos, Nikolaidis e Pitas (2007).

Nishiyama et al. (2007) propuseram um método para o reconhecimento de faces em vídeos com indivíduos em movimento. Conforme os autores, existem dois problemas ocasionados pelo movimento, oclusões e variações de pose/iluminação. Assim, múltiplas câmeras são

utilizadas para aquisição de faces quase frontais para evitar oclusão e de faces de perfil. O casamento hierárquico imagem-conjunto (*Hierarchical Image-Set Matching* – HISM) cria uma distribuição de cada indivíduo pela integração de um conjunto de faces de uma mesma pessoa adquirida por meio de múltiplas câmeras. A partir de uma face detectada, pontos fiduciais são extraídos e a normalização da pose é feita com base no casamento de um modelo ativo de forma (*Active Shape Models* – ASM) (COOTES et al., 1995). Por fim, uma normalização da iluminação é realizada, conforme ilustrado na Figura 2.10. Resultados experimentais utilizando sequências de vídeo contendo 349 pessoas permitiram constatar que o método alcança um desempenho elevado de reconhecimento em comparação com métodos convencionais que utilizam identificação quadro-a-quadro e uma distribuição obtida a partir de uma única câmera. Por outro lado, duas desvantagens podem ser destacadas ao se utilizar modelos ASM: (i) elevado custo computacional para construção do modelo pois uma grande quantidade de amostras é necessária para a convergência durante o treinamento; e (ii) posicionamento inicial da máscara (*shape*) para otimização (*fitting*) do modelo, caso a inicialização não seja realizada próxima à região facial o processo de busca tente a falhar (THAI e TRUONG, 2011).

Figura 2.10 – O fluxo de geração da imagem de face a fim de aliviar a variação de pose e iluminação.



Fonte: Adaptado de Nishiyama et al. (2007).

Fukui e Yamaguchi (2007) introduziram o método *kernel orthogonal mutual subspace* (KOMSM) para o reconhecimento de objetos 3D. KOMSM é um método fundamentado em aparência para classificação de um conjunto

de padrões tais como quadros de um vídeo ou um conjunto de imagens obtidos de um sistema multi-câmera. KOMSM é uma extensão não-linear do método *mutual subspace* (MSM) com o uso de um *kernel* que classifica um conjunto de padrões com base no ângulo canônico Θ entre subespaços de classes lineares. Uma avaliação experimental foi conduzida em uma base de imagens de faces 2D privada de 50 pessoas capturadas sobre 10 diferentes condições de iluminação (a face representa uma região de tamanho 15 a 15 de uma imagem de entrada de tamanho 320 a 240), na qual foi obtida uma acurácia de 97,42%. Embora o método KOMSM seja capaz de lidar com uma variabilidade de padrões atingindo uma alta acurácia, seu desempenho tende a cair significativamente quando as distinções dos padrões possuem uma grande quantidade de estruturas não-lineares, pois nesses casos as distribuições de classe não podem ser representadas por um subespaço linear sem que haja sobreposição, afetando negativamente a classificação.

2.2.2. Abordagens com base em Sequências de Quadros

As abordagens com base em sequências de quadros tratam o problema de agrupamento de faces em vídeos em termos de casamento de conjuntos de múltiplas amostras.

Anantharajah et al. (2015) propuseram um método para solucionar o problema de agrupamento de indivíduos em vídeos, sendo esta uma tarefa útil para busca em vídeos, anotação, recuperação e identificação de elenco. A abordagem proposta, denominada *Local Total Variability Modelling* (Local TVM), é dividida em duas etapas de agrupamento. Na primeira etapa, as faces são agrupadas com respeito à cor e informações locais e espaciais. Na segunda etapa, utiliza-se uma modelagem com base em rastreamento de faces e um agrupamento aglomerativo hierárquico para realizar o agrupamento ao longo do vídeo. As faces são detectadas pelo método de Viola e Jones (2001) e a aparência da face é modelada usando características de covariância (TUZEL, PORIKLI e MEER, 2006). Experimentos foram realizados na base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013) que é subdivida em dois subconjuntos *dev* (*development*) e *eval* (*evaluation*), nos quais foram obtidos os valores

98,88% e 70,90% para as métricas *Average Purity* (P_w) e *Average Coverage* (C_w), respectivamente no subconjunto *dev*, e 98,10% e 77,00% para as métricas *Average Purity* (P_w) e *Average Coverage* (C_w), respectivamente no subconjunto *eval.*

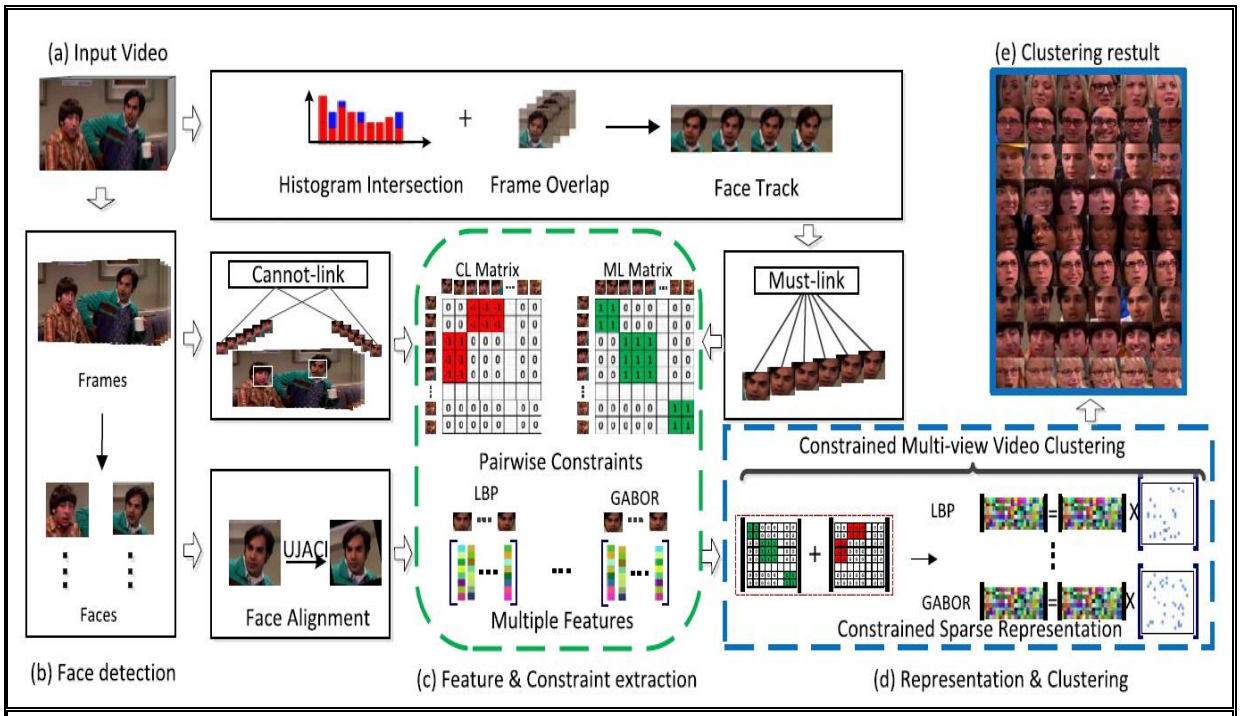
Zhou et al. (2015) propuseram um método de agrupamento de faces denominado, *Multicue Augmented Face Clustering* (McAFC). Tal abordagem é representada por duas características: (i) par de restrições *cannot-link* e *must-link* que podem ser extraídos do conhecimento temporal e espacial dos dados; e (ii) associação de uma série de atributos que pode contribuir para a discriminação entre as faces. Por fim, essas informações são incorporadas a uma técnica de agrupamento por grafos-dirigidos esparsos, de forma que, somente faces de um mesmo indivíduo sejam conectadas. O experimento realizado na base de vídeos *YouTube Faces* (WOLF et al., 2011) obteve o valor de 84,86% para a métrica *Rand Index* (RI), indicando o nível de qualidade do agrupamento realizado pelo método.

Tang et al. (2015) investigaram o problema de agrupamento de faces em vídeos com auxílio das falas dos atores presentes no texto do *script*. Os autores propõem um modelo probabilístico de agrupamento denominado, *Hidden Conditional Random Field* (HCRF). Este modelo incorpora as restrições *must-link* e *cannot-link* com adição de informação externa ao vídeo, como o texto do *script*. O modelo é inicializado via *K-Means* e otimizado com o método *Expectation-Maximization* (EM). As características faciais são extraídas com base na detecção dos olhos e, em seguida, são extraídas características *Discrete Cosine Transform* (DCT) para composição do vetor de características, conforme proposto no trabalho de Bauml et al. (2013). O experimento realizado na base pública de vídeos *Big Bang Theory* – BBT (BAUML et al., 2013) obteve o valor de 69,90% de acurácia de agrupamento, indicando uma significativa complexidade da base de vídeos.

Cao et al. (2015B) estenderam o trabalho anterior (CAO et al., 2015A) focando no cenário de vídeos digitais. Os autores propuseram um novo método de agrupamento denominado, *Constrained Multi-view Video Face Clustering* (CMVFC), com base em restrições *cannot-link* e *must-link* e

modelado na forma de grafo unificado, reforçando as restrições relativas à penalização entre a dissimilaridade *multi-view* entre os diferentes grafos. As faces são inicialmente detectadas e rastreadas, respeitando-se as restrições *cannot-link* e *must-link*. A representação facial é composta de características *Local Binary Patterns* – LBP (OJALA, PIETIKAINEN e HARWOOD, 1996) e Gabor (RIESENHUBER e POGGIO, 1999), conforme ilustrado na Figura 2.11. Experimentos foram realizados em duas bases públicas de vídeos, *Big Bang Theory* – BBT (BAUML et al., 2013) e *YouTube Faces* (WOLF et al., 2011) com valores da métrica de avaliação de agrupamento *Normalized Mutual Information* (NMI) de 92,07% e 60,70%, respectivamente.

Figura 2.11 – Diagrama em blocos da abordagem proposta por Cao et al. (2015B).

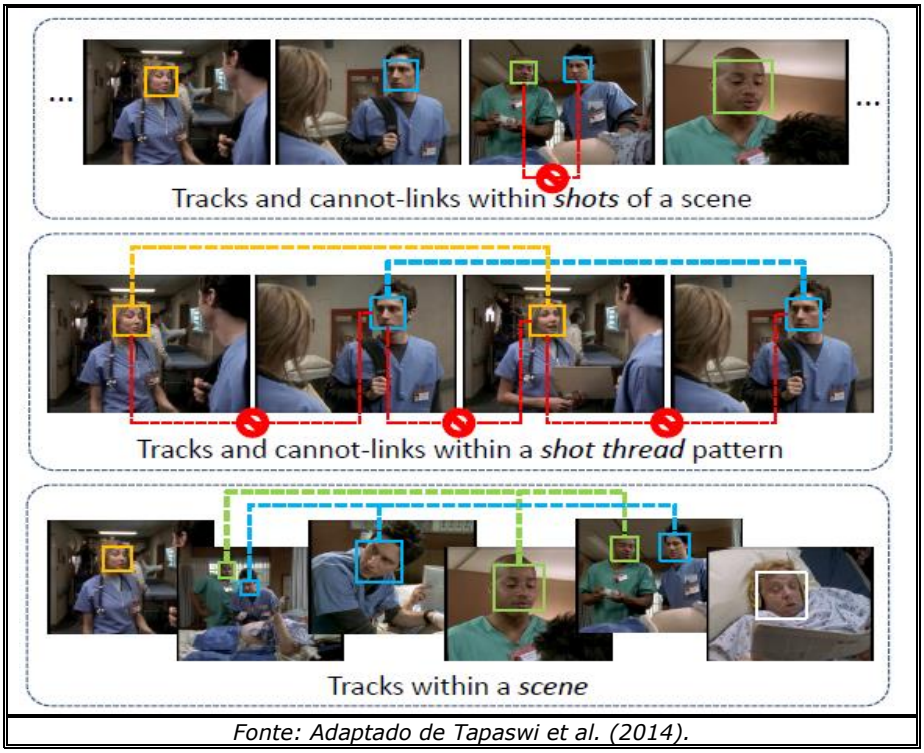


Fonte: Adaptado de Cao et al. (2015B).

O estudo de Tapaswi et al. (2014) objetiva o agrupamento não-supervisionado de faces em materiais extraídos de vídeos, tais como cenas e *shots*. Os autores argumentam sobre os benefícios da utilização de informações sobre cenas e *shots* como fator determinante na qualidade do agrupamento automático de *face tracks*, reduzindo o número de erros. A detecção de cenas e *shots* é realizada de acordo com o método *Displaced Frame Difference* – DFD (YUSOFF, CHRISTMAS e KITTLER, 1998), o qual captura as diferenças de compensação de movimentos entre quadros consecutivos, produzindo picos que identificam os *shots*. O método DFD tem

como função limitar a extensão temporal de uma *face track*. Em outras palavras, todas as detecções em um mesmo *face track* devem pertencer a um mesmo *shot*, conforme ilustrado na Figura 2.12. A representação facial é composta por descritores *Scale Invariant Feature Transform* – SIFT (LOWE, 1999). O algoritmo de agrupamento adotado é o *Hierarchical Agglomerative Clustering* – HAC. Experimentos foram realizados nas bases *Scrubs* (APOSTOLOFF e ZISSERMAN, 2007) e *Buffy* (EVERINGHAM, SIVIC e ZISSERMAN, 2006), nos quais foram obtidos os valores 99,20% e 99,80% para a métrica *Average Purity* (P_w), respectivamente.

Figura 2.12 – Diagrama da abordagem proposta por Tapaswi et al. (2014).

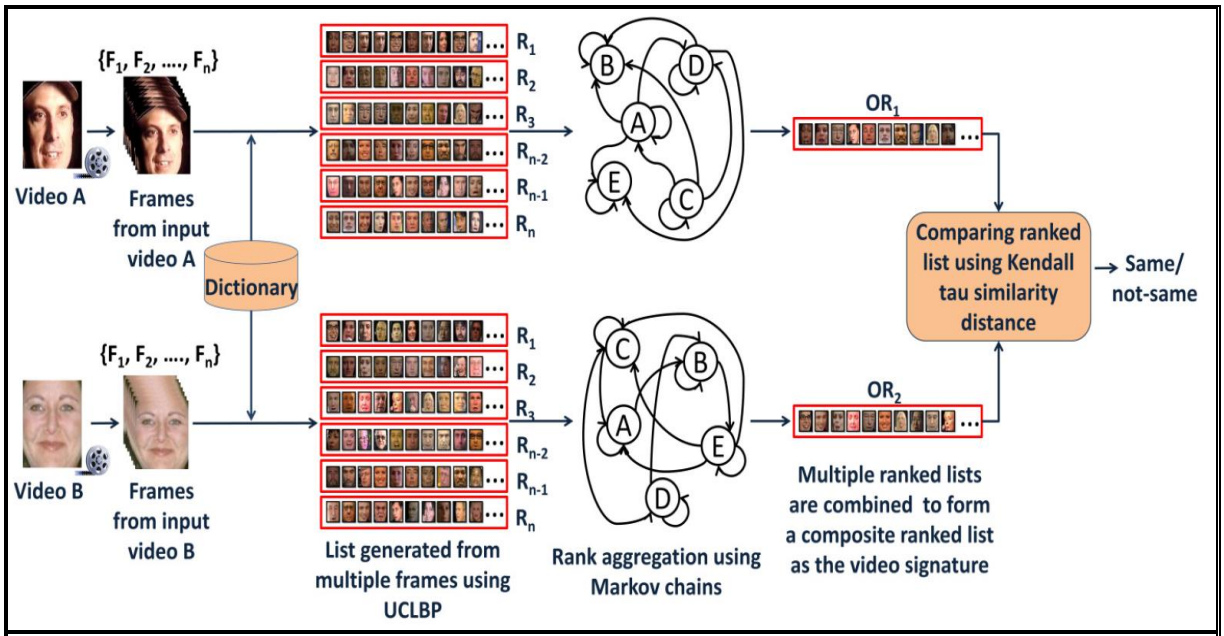


Fonte: Adaptado de Tapaswi et al. (2014).

Bhatt et al. (2014) abordaram o problema da comparação de dois vídeos digitais. Os autores apresentaram um algoritmo de reconhecimento de faces em vídeos que calcula uma assinatura discriminativa do vídeo de entrada como uma lista ordenada de imagens de faces. A assinatura do vídeo incorpora diversas variações intra-pessoais e temporais entre quadros, facilitando a discriminação entre dois vídeos com grandes variações faciais. A comparação é realizada em função da medida de similaridade *Kendall tau* entre as assinaturas discriminativas dos dois vídeos. As variações intra-pessoal e temporal são combinadas por meio de uma cadeia de Markov com base em uma abordagem de ranqueamento por

agregação, conforme ilustrado na Figura 2.13. Experimentos foram realizados em duas bases públicas de vídeos *YouTube Faces* (WOLF et al., 2011) e *Multiple Biometric Grand Challenge* – MBGC (PHILLIPS et al., 2009) com acurácia de 80,70% e 62,20%, respectivamente.

Figura 2.13 – Diagrama em blocos da abordagem proposta por Bhatt et al. (2014) para a comparação de dois vídeos digitais.



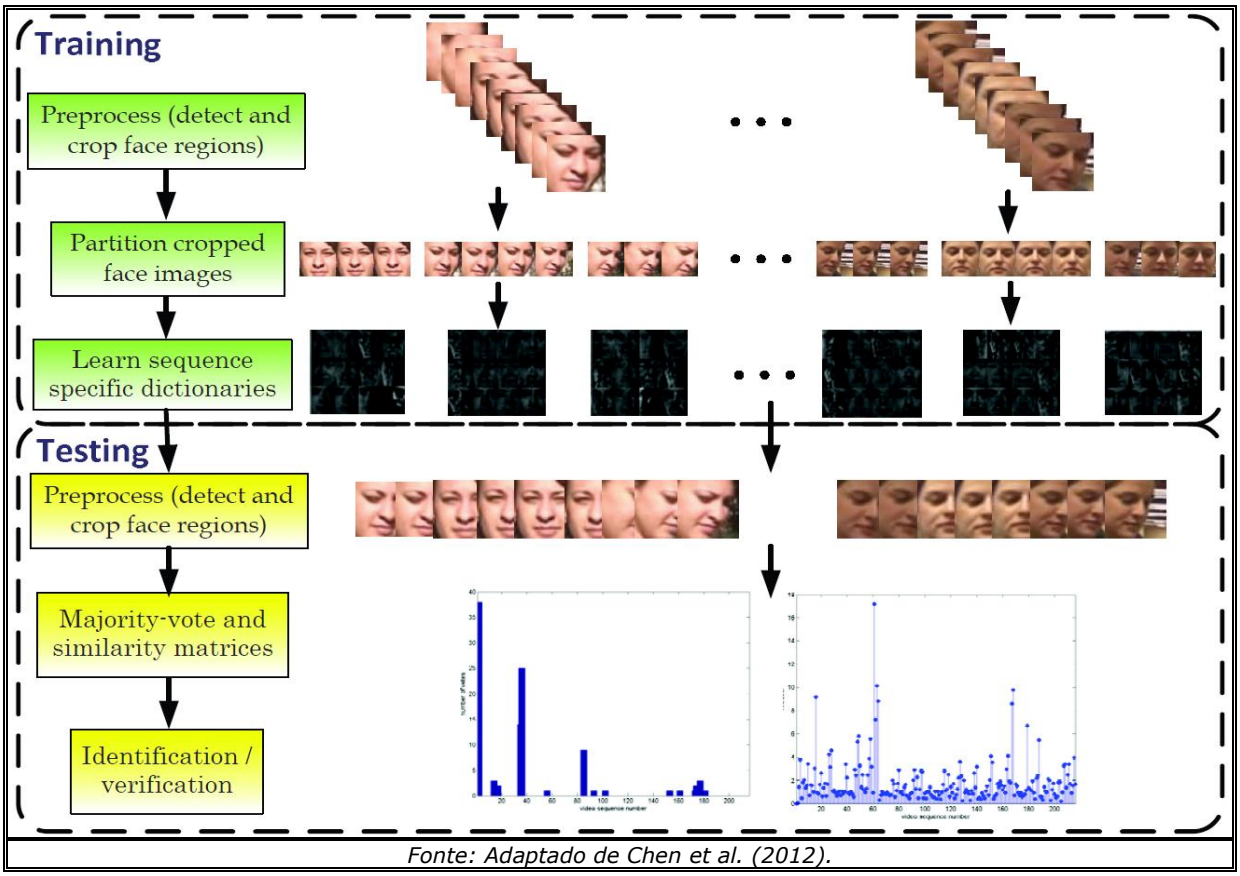
Fonte: Adaptado de Bhatt et al. (2014).

O algoritmo proposto por Mian (2013) combina um conjunto representativo de características SIFT locais extraídas de múltiplas imagens de faces com uma técnica de agrupamento hierárquico e um esquema de votação. As características SIFT permitem robustez à oclusão e à rotação. A similaridade entre um par de faces é mensurada em termos do ângulo e do número de correspondências entre os vetores SIFT. A média ponderada destas medidas fornece o grau de similaridade com que o algoritmo de agrupamento hierárquico opera. O autor escolheu um particionamento manual, especificando a priori o número de grupos finais, assim, o cálculo do ponto de corte que determina a composição final dos grupos não é feita de maneira automática. Cada um dos grupos contém faces com aparências semelhantes, de modo que múltiplos conjuntos com expressões e poses semelhantes correspondam a uma mesma pessoa. Um processo de votação realizado durante a comparação de faces é utilizado para selecionar o conjunto representativo de características. Uma acurácia de 99,60% foi alcançada na base de dados *Honda UCSD* (LEE et al., 2005) com esta

abordagem.

Chen et al. (2012) elaboraram uma abordagem generativa para o problema de reconhecimento de faces em vídeo, na qual uma sequência de vídeos é particionada em subsequências e dicionários específicos de cada subsequência foram aprendidos. Os quadros de cada vídeo de pesquisa são projetados no espaço de cada dicionário específico de sequência e o resultado da projeção é utilizado para reconhecimento. Por fim, determina-se uma representação esparsa multivariada que simultaneamente considera a correlação e o acoplamento de informações entre os quadros, conforme ilustrado na Figura 2.14. Foi realizado um experimento na base intitulada *Multiple Biometric Grand Challenge* – MBGC (PHILLIPS et al., 2009) contendo 770 sequências de vídeo de 146 diferentes pessoas com uma acurácia de 59,00%. Esta abordagem requer a criação de vários dicionários específicos de sequência para variações de pose e iluminação, o que aumenta o custo computacional.

Figura 2.14 – Fluxograma da abordagem proposta por Chen et al. (2012).

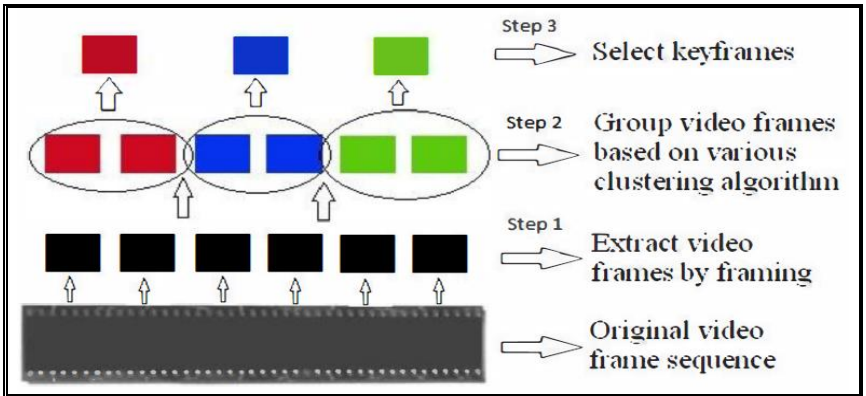


Fonte: Adaptado de Chen et al. (2012).

Sony et al. (2011) propuseram uma técnica de sumarização para

reunir apenas quadros de interesse em um vídeo que consiste na remoção de quadros redundantes e na manutenção de um número fixo de quadros definido pelo usuário. O método funciona de forma que quadros visualmente semelhantes são agrupados de acordo com a medida de distância euclidiana. Quando os grupos são formados, frações de quadros que apresentam maior similaridade entre si formam uma sequência que compõe a saída desejada, conforme ilustrado na Figura 2.15. Os autores afirmam que o vídeo sumarizado preserva os quadros mais relevantes do vídeo de entrada e que a continuidade do vídeo é mantida. Por outro lado, para um vídeo com um grande número de descontinuidades a saída poderá apresentar resultados imprecisos, uma vez que a técnica proposta assume a continuidade do vídeo. Um experimento realizado em um vídeo com 1351 quadros apresentou uma taxa de sumarização (razão entre o número de quadros sumarizados e o número de quadros de entrada) de 0,521 resultando num total de 704 quadros. Esta abordagem apesar de ser considerada genérica, pode ser aplicada para o contexto de faces (e.g., sumarização de vídeos de segurança).

Figura 2.15 – Diagrama da abordagem proposta por Sony et al. (2011).



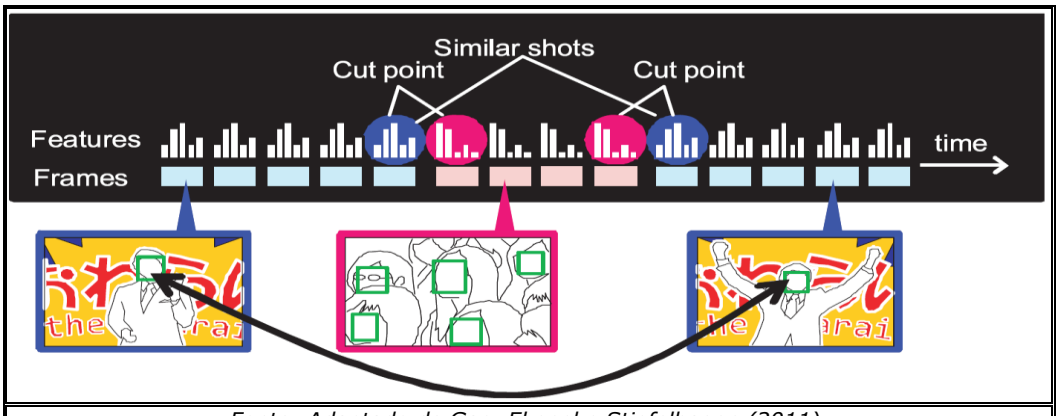
Fonte: Adaptado de Sony et al. (2011).

Gao, Ekenel e Stiefelhagen (2011) desenvolveram um sistema que detecta automaticamente uma lista de pessoas específicas, tais como apresentadores âncora ou políticos em vídeos de noticiários. O método é baseado na indexação de faces detectadas e no reconhecimento por meio de uma lista pré-definida de faces candidatas. O sistema identifica certa pessoa verificando se a face está presente na lista alvo, atribuindo então uma identidade para a mesma. Experimentos realizados mostraram que o

sistema obteve uma taxa média de precisão de valor 92,60%, com uma taxa média de cobertura de valor 75,40%. Um possível incremento no referido trabalho seria a exploração de técnicas mais robustas a variações de pose, iluminação, expressões faciais, entre outras, presentes no mundo real de identificação de faces.

Yamamoto, Yamaguchi e Aoki (2010) elaboraram uma abordagem para catalogar cenas de um vídeo em função da presença de atores. O objetivo principal é a obtenção de um tempo de espera tolerável após a gravação, definido como menos de 3 minutos por hora de vídeo. Os autores ressaltaram que métodos de reconhecimento de faces convencionais que utilizam técnicas de agrupamento podem obter bons resultados, no entanto exigem um tempo de processamento considerável. De maneira a reduzir o tempo de processamento foram utilizadas semelhanças entre tomadas nas quais os personagens aparecem para estimar as faces dos atores em quadros correspondentes, em vez de calcular a distância entre cada característica facial, conforme ilustrado na Figura 2.16. Experimentos realizados mostram que o arcabouço proposto pelos autores diminuiu em 6% o tempo de processamento em comparação com métodos de reconhecimento de faces convencionais que utilizam características faciais, obtendo um tempo médio de espera de 2,8 minutos.

Figura 2.16 – Uma ilustração conceitual da abordagem proposta por Gao, Ekenel e Stiefelhagen (2011).



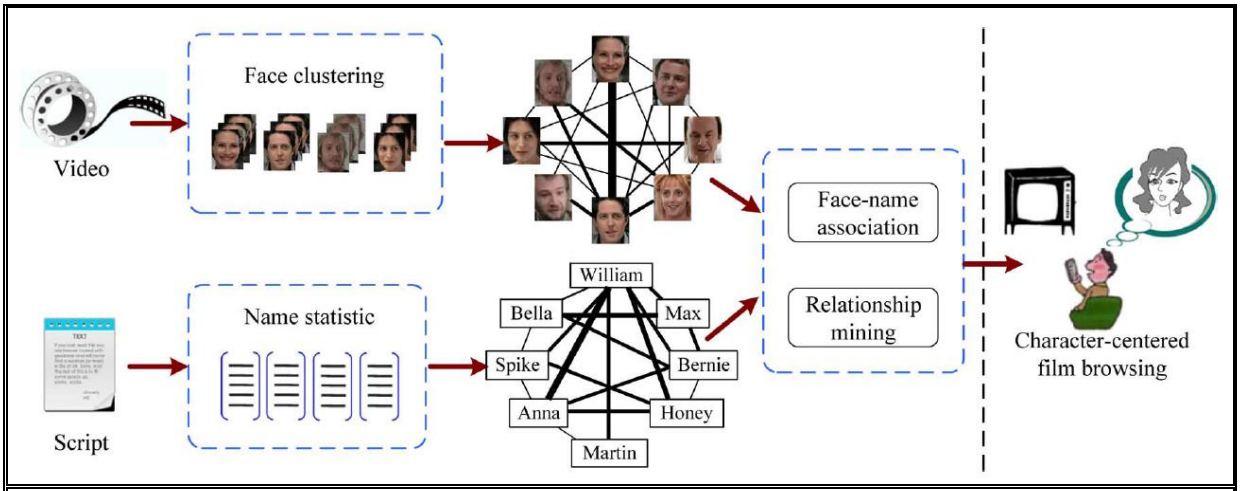
Fonte: Adaptado de Gao, Ekenel e Stiefelhagen (2011).

Turaga, Veeraraghavan e Chellappa (2009) abordaram o problema de indexação de vídeos a partir da atividade dos objetos, ou seja, os quadros de uma sequência de vídeo podem ser agrupados de modo que cada grupo represente uma atividade. Para tratar o problema, as atividades dos objetos

são descritas como uma cascata de sistemas dinâmicos que aumentam o poder expressivo do modelo representativo das atividades, aproveitando vantagens computacionais inerentes aos modelos dinâmicos lineares (*Linear Dynamical Systems Model* – CLDS). Em seguida, métodos foram derivados para incorporar invariância ao ponto de vista e à taxa de execução da atividade nos modelos. Em um experimento realizado em um vídeo de 10 minutos, composto de 4 diferentes padrões de atividade, apenas 3 sequências foram classificadas erroneamente de um total de 24 sequências que correspondem a atividades com significado semântico.

Zhang et al. (2009) investigaram o problema de identificação de personagens em longas-metragens a partir do roteiro do filme, permitindo ao usuário usar o nome do personagem para procurar vídeos relacionados. A abordagem proposta compreende correspondências locais entre uma face e um dos nomes extraídos da transcrição temporal local do vídeo. Um método de grafos correspondentes é utilizado para construir associações face-nome entre uma rede de afinidades de faces e uma rede de afinidades de nomes que são, respectivamente, derivados de seus próprios domínios (vídeo e roteiro do filme), conforme ilustrado na Figura 2.17. Experimentos realizados no trabalho de Zhang et al. (2009) apresentaram uma acurácia média de associação face-nome de 83,28%. Possíveis incrementos ao referido trabalho são a integração de informações de gênero e contexto para refinar o resultado da busca e a geração de um *clip* do vídeo contendo um grupo de personagens previamente selecionados pelo usuário.

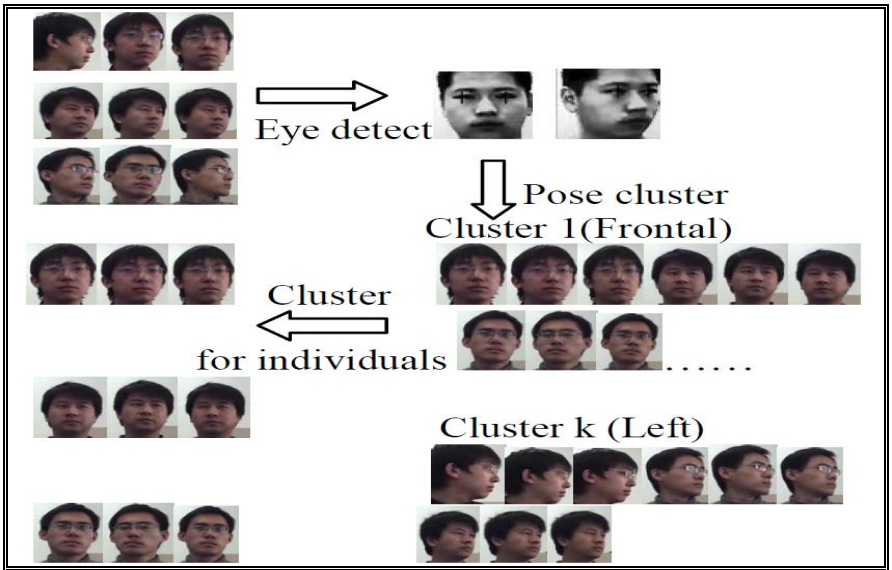
Figura 2.17 – Esquema de identificação de personagens proposta por Zhang et. al. (2009).



Fonte: Adaptado de Zhang et. al. (2009).

Huang, Wang e Shao (2008) ressaltaram que para o problema de agrupamentos de faces multi-pose, a similaridade entre as faces de diferentes indivíduos com poses semelhantes é geralmente maior do que a similaridade entre faces multi-pose do mesmo indivíduo, e que isto pode exercer um impacto negativo sobre o resultado final do agrupamento. Para contornar este problema, os autores adotaram a estratégia de primeiro efetuar o agrupamento por pose, e, em seguida, realizar o agrupamento de diferentes pessoas, conforme ilustrado na Figura 2.18. Para a determinação da pose, foram utilizadas as coordenadas dos olhos. Uma das limitações do trabalho está no cenário no qual existem situações com grandes variações de pose, como perfil lateral completo (*full-profile*). Assim, a estratégia proposta pode não ser satisfatória dado que apenas um dos olhos estará visível. Experimentos realizados em dois vídeos apresentaram taxas de classificação de 81% (sem agrupamento por pose) e de 90,1% (com agrupamento por pose), respectivamente, comprovando a eficácia da estratégia proposta.

Figura 2.18 – Diagrama da abordagem proposta por Huang, Wang e Shao (2008).



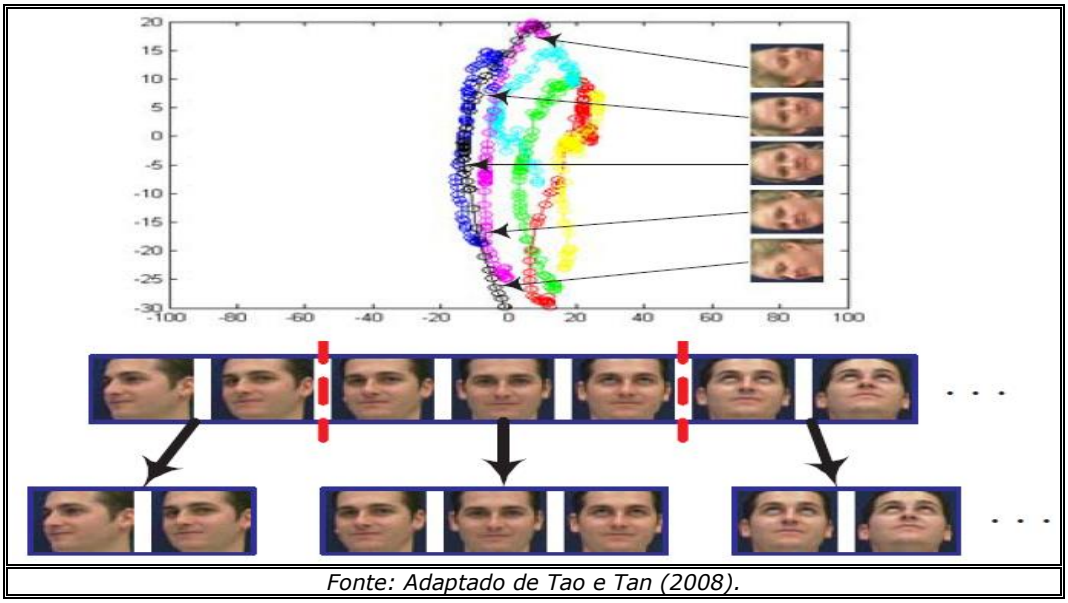
Fonte: Adaptado de Huang, Wang e Shao (2008).

Kim et al. (2008) abordaram o problema do rastreamento e reconhecimento de faces em vídeos de baixa qualidade, em especial de alguns vídeos postados no *YouTube*. Faces são rastreadas usando-se um rastreador que adaptativamente constrói uma combinação de modelos generativos e discriminativos que refletem mudanças na aparência, situação típica em vídeos. A parte generativa confronta a face com um espaço

genérico de poses, enquanto a parte discriminativa garante a rejeição de faces com problemas de alinhamento, melhorando a robustez do rastreador quanto a mudanças abruptas e oclusões. A identidade da face rastreada é feita pela fusão de características discriminantes à pose e à pessoa ao longo da duração da sequência de vídeo. Experimentos realizados indicaram valores da métrica *Average Purity* (P_w) de 100% na base de dados *Honda/UCSD* (LEE et al., 2005) e 70% em um conjunto do *YouTube* com 35 celebridades e 1500 sequências.

Tao e Tan (2008) propuseram uma abordagem para o agrupamento de faces humanas em vídeos, aplicada à identificação do elenco principal (*cast list*). Em cada tomada do vídeo, as faces detectadas são, primeiramente, associadas para formar sequências. Ao invés de comparar as sequências de faces diretamente, a técnica proposta particiona as mesmas em subsequências contendo poses similares, de modo a facilitar a comparação, conforme ilustrado na Figura 2.19. As subsequências de faces são agrupadas por particionamento de grafos, com o auxílio de uma matriz de similaridade. Apesar dos autores considerarem sua abordagem promissora, tal afirmação pode ser criticada devido à fraca avaliação experimental realizada, dado que foram utilizados apenas dois vídeos e nenhuma métrica de avaliação da qualidade do agrupamento foi levada em consideração.

Figura 2.19 – Exemplo de particionamento de uma sequência de faces em três subsequências do trabalho de Tao e Tan (2008).



Fonte: Adaptado de Tao e Tan (2008).

Ramanan, Baker e Kakade (2007) introduziram um método semi-supervisionado para a construção de grandes conjuntos de dados de faces rotuladas por meio da indexação de vídeos. A técnica proposta permite extrair modelos de histogramas de cores da face, cabelo e torso, que são utilizados por um conjunto de detectores. Quadros vizinhos são agrupados em torno das detecções por meio da adição de restrições dinâmicas e pela aplicação de um algoritmo de agrupamento aglomerativo nas regiões acima mencionadas (face, cabelo e torso). Em seguida, grupos com a mesma identidade são mesclados. Experimentos realizados em 22 episódios do seriado de televisão *Friends* demonstraram que o método proposto se apresenta robusto quanto a variações de idade, ganho de peso e mudança de cabelo dos personagens.

2.3. Considerações sobre os Trabalhos Analisados

Os estudos analisados anteriormente foram agrupados por categoria e por ordem cronológica decrescente, da mesma forma que no Quadro 2.1. Nesse quadro, a coluna *Categoria* indica qual a propriedade do vídeo foi explorada pela abordagem proposta, a coluna *Característica* indica que características foram utilizadas para construção da representação facial, a coluna *Técnica de Agrupamento* especifica as técnicas desenvolvidas ou empregadas para a composição do sistema, a coluna *Base de Imagens* explicita as bases de faces adotadas para verificar a acurácia ou agrupamento do sistema proposto, e finalmente, a coluna *Taxa de Avaliação* especifica os valores dos resultados experimentais obtidos.

Inúmeros problemas associados à abordagem temporal a partir de sequências de faces ainda permanecem sem solução (CHEN et al., 2012). Um tópico em aberto é se a presença ou não de movimentos faciais com durações mais longas permitiriam o reconhecimento facial mais confiável ou mais preciso (MIAN, 2013). Outro problema diz respeito à complexidade amostral que o reconhecimento de faces em vídeo pode implicar devido ao número elevado de quadros extraídos (BHATT et al., 2014).

Constata-se que apesar dos grandes avanços alcançados na área de agrupamento de faces em vídeos, ainda há muito trabalho a ser feito. De

acordo com a revisão da literatura, os sistemas destinados ao agrupamento de faces obtiveram elevadas taxas de reconhecimento quando imagens de faces foram adquiridas em condições controladas, como por exemplo, quando utilizando as bases Honda UCSD (LEE et al., 2005) e CMU Mobo (GROSS e SHI, 2001). Todavia, tal situação está muito distante da realidade, na qual existe uma necessidade crescente de sistemas cada vez mais robustos, i.e., mais tolerantes a variações de pose, iluminação, oclusão e expressão facial, conforme amostras encontradas nas bases *YouTube Celebrities* (KIM et al., 2008) e *YouTube Faces* (WOLF et al., 2011).

Deste modo, outro ponto desafiador está na obtenção de uma base de dados que incorpore grande quantidade de pessoas, oclusões, ruído e artefatos de compressão, juntamente com variações na pose, iluminação e expressão. Tais aspectos são necessários para avaliar o desempenho de sistemas em ambientes não controlados do mundo real (KIM et al., 2008, WOLF et al., 2011).

Percebe-se também que poucas pesquisas têm sido conduzidas na adaptação de rastreadores de faces robustos para encontrar faces em ambientes não controlados do mundo real. Rastreadores de faces são geralmente mais eficientes do que os detectores já que não realizam uma pesquisa completa sobre todos os quadros, e podem, potencialmente, melhorar o desempenho geral do sistema.

Durante o desenvolvimento da abordagem proposta, foi verificado que a característica SURF (BAY, TUYTELAARS e VAN GOOL, 2006), apesar de apresentar comportamento similar, apresentou melhores resultados do que a característica SIFT. Experimentos comparativos apresentados no trabalho de Juan e Gwon (2009) evidenciam que o SIFT deve ser preterido ao SURF em relação a desempenho na geração dos descritores (cerca de 3 vezes mais lento) e na acurácia de 78,1% contra 85,7% do SURF na base de imagens Caltech (CALTECH FACE DATABASE, 2010).

No tocante às características extraídas, PCA e LBP são as mais presentes. A métrica da avaliação mais adotada foi a *Average Purity* (P_w), seguida das métricas *F-Measure*, *Precision* e *Recall*. No que diz respeito à

tarefa de agrupamento ou reconhecimento, a categoria de *técnicas baseadas em sequência de quadros* é a mais presente.

Em relação à estratégia de comparação de faces, foi verificado que a utilização de distâncias como a Euclidiana e Cosseno, não são compatíveis com os descritores SURF. Utilizou-se o algoritmo *Fast Approximate Nearest Neighbors* – FANN (MUJA e LOEW, 2009), para determinar o grau de similaridade entre duas faces em função da correspondência entre descritores e inspirado na abordagem proposta por Antonopoulos, Nikolaidis e Pitas (2007). Adicionalmente, não existe uma métrica de similaridade que seja unânime na literatura, da mesma forma que não há consenso nas métricas de avaliação de agrupamento mais comumente adotadas para mensurar a qualidade final dos grupos gerados.

Desta forma, pode-se concluir que ainda não foi desenvolvido um sistema para o agrupamento de faces em vídeo que seja robusto a uma variabilidade de condições, tais como iluminação, pose, expressão facial e oclusão de partes da face. Assim, uma opção para alcançar este objetivo seria criar um sistema híbrido que explorasse o melhor de cada técnica.

2.4. Considerações Finais

Neste capítulo, foram apresentados métodos, métricas e técnicas que serviram de base para a consecução da pesquisa ora desenvolvida, um sistema de agrupamento de faces em vídeos robusto a variações de iluminação, expressão facial e pose, com o intuito de facilitar a organização e extração de conteúdo relevantes para os usuários.

Do resumo apresentado no Quadro 2.1 e da exposição anterior, evidencia-se o fato de que a área de agrupamento de faces em vídeo tem despertado interesse de muitos pesquisadores nos últimos anos (OTTO, WANG e JAIN, 2016; SCHROFF, KALENICHENKO e PHILBIN, 2015; CAO et al., 2015A; CAO et al., 2015B; ANANTHARAJAH et al., 2015; ZHOU et al., 2015; TANG et al., 2015; TAPASWI et al., 2014; BHATT et al., 2014). Este fato é reforçado pela crescente demanda por métodos de organização automática e extração de informações relevantes, devido ao

Quadro 2.1 – Resumo dos trabalhos analisados.

Categoria	Artigo	Característica	Técnica de Agrupamento	Base de Imagens	Taxa de Avaliação
Com base em Conjuntos de Quadros	Otto, Wang e Jain (2016)	Pontos Fiduciais	*Rank-Order Clustering*	LFW YouTube Faces	F-Measure: 87,00% F-Measure: 71,00%
	Schroff, Kalenichenko e Philbin (2015)	LMNN	Agrupamento aglomerativo	LFW YouTube Faces	ACC: 99,63% ACC: 95,12%
	Cao et al. (2015A)	Gabor e LBP	*Diversity-induced Multi-view Subspace Clustering*	Yale Yale B ORL	ARI: 53,50% ARI: 45,30% ARI: 80,20%
	Anoop et al. (2012)	Gabor	*Spectral Clustering*	Sitcom YouTube Celebrities	P_w: 90,76% P_w: 80,35%
	Cui et al. (2012)	Pontos Fiduciais	Alinhamento de conjunto de quadros	Honda UCSD CMU Mobo YouTube Celebrities	P_w: 98,90% P_w: 95,00% P_w: 74,60%
	Wolf et al. (2011)	CSLBP, FPLBP e LBP	Similaridade entre conjuntos	YouTube Faces	ACC: 72,60%
	Hu et al. (2011)	LBP	*Sparse Aprroximated Nearest Point*	Honda UCSD CMU Mobo YouTube Celebrities	P_w: 92,31% P_w: 97,08% P_w: 65,03%
	Harandi et al. (2011)	LDA	*Grassmannian manifolds*	CMU PIE BANCA CMU Mobo	ACC: 75,80% ACC: 68,73% ACC: 89,92%
	Wang et al. (2008)	PCA	Distância entre centroides	Honda UCSD CMU Mobo	ACC: 96,90% ACC: 93,60%
	Antonopoulos, Nikolaidis e Pitas (2007)	SIFT	Agrupamento hierárquico	Privada	F-Measure: 87,60%
	Nishiyama et al. (2007)	ASM, Pontos Fiduciais	Casamento hierárquico imagem-conjunto	Privada	ACC: 97,40%
	Foucher e Gagnon (2007)	PCA	Agrupamento com base em cascata de classificadores	Privada	ACC: 25,00%
	Fukui e Yamaguchi (2007)	PCA	*Kernel orthogonal mutual subspace*	Privada	ACC: 97,42%

Quadro 2.1 – Resumo dos trabalhos analisados (continuação).

Categoria	Artigo	Característica	Técnica de Agrupamento	Base de Imagens	Taxa de Avaliação
Com base em Sequências de Quadros	Anantharajah et al. (2015)	Local TVM	Agrupamento hierárquico	SAIVT-Bnews	P_w: 98,88% C_w: 70,90%
	Zhou et al. (2015)	PCA	*Multicue Augmented Face Clustering*	YouTube Faces	RI: 84,86%
	Tang et al. (2015)	DCT	*Hidden Conditional Random Field*	BBT	ACC: 69,90%
	Cao et al. (2015B)	Gabor e LBP	Constrained Multi-View Video Face Clustering	BBT YouTube Faces	NMI: 92,07% NMI: 60,70%
	Tapaswi et al. (2014)	SIFT	Agrupamento aglomerativo hierárquico	Scrubs Buffy	P_w: 99,20% P_w: 99,80%
	Bhatt et al. (2014)	UCLBP	Agrupamento com ranqueamento	YouTube Faces MBGC	ACC: 80,70% ACC: 62,20%
	Mian (2013)	SIFT	Agrupamento hierárquico	Honda UCSD	ACC: 99,60%
	Chen et al. (2012)	PCA	Dicionários de vídeos	MBGC	ACC: 59,00%
	Sony et al. (2011)	Pontos Fiduciais	Agrupamento por distância euclidiana	Privada	N/A
	Gao, Ekenel e Stiefelhagen (2011)	DCT	Agrupamento por *face tracks*	Privada	P: 92,60% R: 75,40%
	Yamamoto, Yamaguchi e Aoki (2010)	Histograma de Cor	Agrupamento por *shot similarity*	Privada	ARI: 90,0%
	Turaga, Veeraraghavan e Chellappa (2009)	PCA, LDA	Agrupamento com base na atividade dos objetos	UMD USF INRIA	Número de grupos finais
	Zhang et al. (2009)	SIFT, LLE	Agrupamento por casamento de grafos	Privada	ACC:83,30%
	Kim et al. (2008)	LDA	Restrições visuais utilizando modelos generativos	Honda UCSD YouTube Celebrities	P_w: 100,00% P_w: 70,00%
	Tao e Tan (2008)	PCA	Agrupamento por propagação de restrições	Privada	N/A
	Huang, Wang e Shao (2008)	PCA	Agrupamento com base em pose	Privada	ACC: 90,10%
	Ramanan, Baker e Kakade (2007)	Histograma de Cor, PCA	Agrupamento por indexação de indivíduos	Privada	N/A

aumento significativo da quantidade de vídeos digitais por usuários de câmeras digitais e dispositivos móveis multimídia, as quais, em sua grande maioria, contêm pessoas como principal tema (YOUTUBE ADVERTISE, 2014).

No próximo capítulo, apresenta-se, de forma detalhada, a abordagem proposta para a solução do problema objeto de estudo, incluindo-se a arquitetura geral, o fluxo de processamento e o funcionamento de cada módulo. Adicionalmente, são apresentados alguns detalhes de implementação e de organização interna da aplicação desenvolvida para validar a abordagem proposta.

Capítulo 3

Abordagem Proposta

Neste capítulo, é apresentada a abordagem proposta para a solução do problema de agrupamento de faces em vídeos. Com base nos estudos apresentados no Capítulo 2, foi desenvolvida uma abordagem original, fundamentada em técnicas e métodos do estado-da-arte. Os detalhes da arquitetura, dos módulos, das técnicas e dos algoritmos que compõem a abordagem proposta serão apresentados a seguir.

A abordagem proposta tem como principal diferencial a agregação de módulos para atenuar os efeitos da queda de desempenho do agrupamento, normalmente associada a variações de iluminação, expressões faciais e pose, assim como adota uma colaboração entre técnicas de detecção, rastreamento e agrupamento, juntamente com informações espaço-temporais que objetivam aumentar o desempenho e a qualidade do sistema que implementa a abordagem.

Desta forma, apresenta-se uma visão detalhada da abordagem proposta para a solução do problema objeto de estudo, incluindo-se a arquitetura geral do sistema, seguida do fluxo de processamento de cada fase do processo e o funcionamento de cada módulo da técnica elaborada.

A abordagem proposta pode ser dividida em três módulos: (i) Preparação de Conteúdo; (ii) Seleção de Conteúdo; e (iii) Organização de Conteúdo. Mais detalhes acerca de cada um dos módulos da abordagem proposta, assim como seus componentes, são explicitados nas seções seguintes.

3.1. Visão Geral da Arquitetura Proposta

A arquitetura macro da técnica desenvolvida foi inspirada em abordagens

encontradas na revisão bibliográfica (e.g., BHATT et al., 2014; MIAN, 2013; HARANDI et al., 2011; TAO e TAN, 2008; FOUCHER e GAGNON, 2007), sendo ilustrada na Figura 3.1 e descrita sucintamente como segue:

(i) O vídeo de entrada é submetido, primeiramente, ao módulo de preparação de conteúdo, no qual os quadros do vídeo são extraídos e submetidos a duas etapas de *pré-processamento*, a saber: (a) filtragem homomórfica; e (b) equalização de histograma, originando quadros pré-processados. Em seguida, um processo de detecção de *shots* e segmentação de cenas é realizado para auxiliar na etapa de rastreamento de faces;

(ii) Os quadros pré-processados são submetidos ao módulo de seleção de conteúdo, no qual três etapas de processamento são realizadas, a saber: (a) detecção de faces: responsável por localizar todas as faces contidas na imagem; (b) rastreamento de faces: responsável por colaborar com a detecção, rastreando as faces previamente detectadas e minimizando a execução do detector, em virtude de seu alto custo de processamento, de modo a originar trajetórias de faces (*face tracklets*); e (c) extração de características: descritores SURF são extraídos para cada uma das faces rastreadas, originando suas respectivas representações faciais;

(iii) Por fim, o módulo de organização de conteúdo é responsável pela geração das listas de grupos de faces de saída. Este módulo compreende três etapas de processamento, a saber: (i) agrupamento de *face tracklets*: por meio das faces representativas dos *face tracklets,* uma etapa preliminar de agrupamento é realizada; (ii) similaridade temporal: ordenação temporal dos grupos de faces originados na etapa anterior; e (iii) reagrupamento espacial: uma segunda etapa de agrupamento é realizada, com base na informação espacial dos *face tracklets* e da matriz de similaridade das representações faciais, para a geração final dos grupos de faces semelhantes.

Figura 3.1 – Arquitetura macro da abordagem proposta.

O propósito da presente pesquisa é a definição de uma abordagem para o agrupamento de faces em vídeos digitais, usando, dentre as técnicas analisadas, os algoritmos que atingiram melhores resultados.

Tomando-se como base artigos relevantes da área de agrupamento de faces revisados no capítulo anterior, algumas técnicas da área de Processamento Digital de Imagens (PDI) e Visão Computacional (VC) foram estudadas e implementadas, com o intuito de serem usadas nos módulos de pré-processamento e no cerne do processamento da abordagem proposta. Cada técnica implementada é apresentada nas seções que seguem.

Os resultados apresentados neste capítulo são resultantes da utilização de técnicas em cujos experimentos foram verificados os melhores resultados, considerando-se cada uma das etapas da abordagem proposta. Tais experimentos são apresentados e discutidos no próximo capítulo. A seguir, cada etapa do fluxo de execução mostrado na Figura 3.1 é detalhada e exemplos de imagens processadas são apresentadas.

3.2. Preparação de Conteúdo

O módulo inicial da abordagem proposta é responsável por realizar a extração dos quadros de vídeo de entrada, uma vez que o processamento é efetuado em cada quadro do vídeo de entrada, conforme ilustrado na Figura 3.2. Após a extração de um quadro do vídeo, a imagem é submetida a um processo de correção proposto por Moura, Gomes e Carvalho (2013).

Esse processo é baseado em dois algoritmos de PDI, utilizados na seguinte ordem: (i) *filtragem homomórfica* (GONZALEZ e WOODS, 2010), visando a aprimorar a qualidade da imagem por meio da compressão da faixa dinâmica de brilho, simultaneamente à expansão do contraste; e (ii) *equalização de histograma* (GONZALEZ e WOODS, 2010), que visa a realçar o contraste da imagem. As imagens de saída deste módulo, denominadas

quadros pré-processados, servem de entrada para a próxima etapa da abordagem proposta, responsável pela detecção/rastreamento de faces e extração de características faciais. O Apêndice C contém um detalhamento dos algoritmos de filtragem homomórfica e equalização de histograma.

Figura 3.2 – Visão detalhada do módulo de preparação de conteúdo.

3.2.1. Detecção de Shots e Segmentação de Cenas

Após a etapa de pré-processamento, um processo de detecção de *shots* e segmentação de cenas é realizado por meio da ferramenta desenvolvida por Apostolidis e Mezaris (2014) para a segmentação automática temporal de vídeos em *shots* e cenas.

Esta ferramenta auxilia a etapa de rastreamento de faces, pois determina o tempo de duração de uma cena, que corresponde ao tamanho do *face tracklet*, bem como o instante de tempo em que uma cena inicia e finaliza, indicando o momento em que o rastreador de faces deverá ser inicializado ou reinicializado.

Shots são sequências de quadros consecutivos capturados sem interrupção por uma única câmera. A transição entre dois *shots* sucessivos do vídeo pode ser abrupta, quando um quadro pertence a um *shot* e o quadro seguinte pertence ao próximo *shot*, ou gradual, quando dois *shots* são combinados usando efeitos cromáticos, espaciais ou espaço-cromáticos (e.g., *fade in/out*, *dissolve*, *wipe*), que gradualmente substituem um *shot* pelo outro.

Tal ferramenta é capaz de detectar ambos os tipos de transições com base na semelhança visual de quadros vizinhos no vídeo.

A eficiência descritiva de dois descritores, um local (SURF) e outro global (HSV), são exploradas para avaliar a similaridade entre quadros vizinhos. Especificamente, transições abruptas são inicialmente detectadas entre quadros de vídeo sucessivos nos quais ocorre uma mudança brusca no conteúdo visual, que se expressa por um valor muito baixo de similaridade entre os mesmos.

Os valores de similaridade calculados são analisados para a identificação de sequências de quadros nas quais uma mudança progressiva do conteúdo visual ocorra, uma transição gradual. Por fim, uma etapa de pós-processamento é realizada com o objetivo de identificar valores de similaridade discrepantes, devidos ao movimento do objeto ou da câmera e iluminação de *flash*.

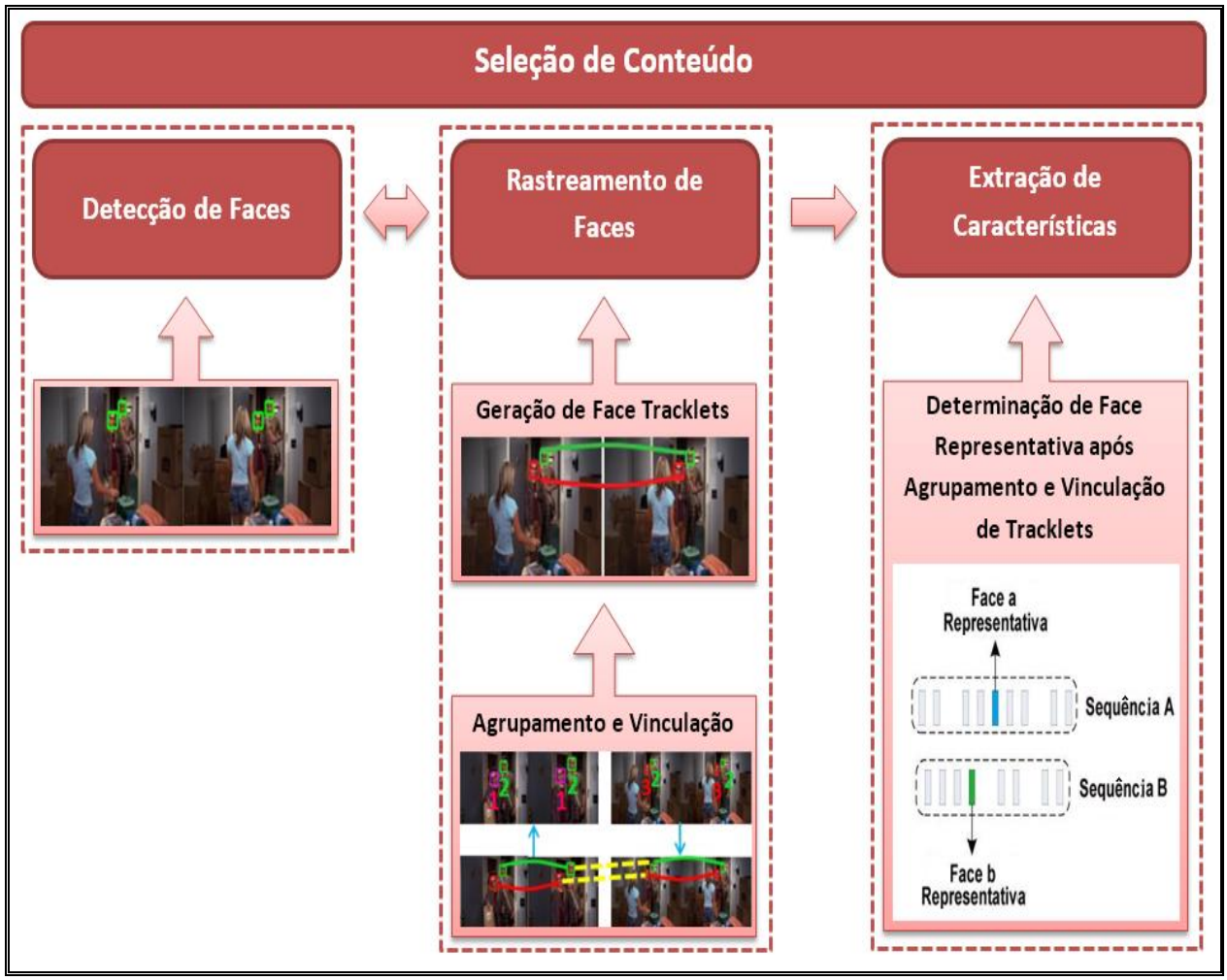
Cenas são fragmentos temporais de nível superior que correspondem às partes mais importantes do conteúdo do vídeo. Elas são formadas ao se agrupar os *shots* detectados em fragmentos temporais de vídeo semanticamente coerentes. No tocante ao tempo de processamento, a computação paralela, por meio de *multi-threads*, é utilizada para fazer múltiplas vezes toda a análise do vídeo, sendo mais rápida do que o processamento em tempo real.

Os experimentos apresentados no trabalho de Apostolidis e Mezaris (2014) demostraram que a ferramenta proposta por eles alcança alta precisão nas tarefas de detecção de *shots* e segmentação de cenas, enquanto é capaz de realizar a análise de um vídeo em tempo mais rápido do que o real.

3.3. Seleção de Conteúdo

O segundo módulo da abordagem proposta nesta tese é responsável por um processamento que engloba três etapas, a saber: (i) detecção de faces; (ii) rastreamento de faces; e (iii) extração de características, conforme ilustrado na Figura 3.3.

Figura 3.3 – Visão detalhada do módulo de seleção de conteúdo.



3.3.1. Detecção de Faces

Considerando que o foco principal desta Tese são faces humanas, a etapa inicial do módulo de *Seleção de Conteúdo* consiste na detecção de faces, bem como na correção da orientação das faces detectadas, pela determinação e localização da posição dos olhos. Este procedimento gera faces normalizadas, com o objetivo de facilitar posterior comparação.

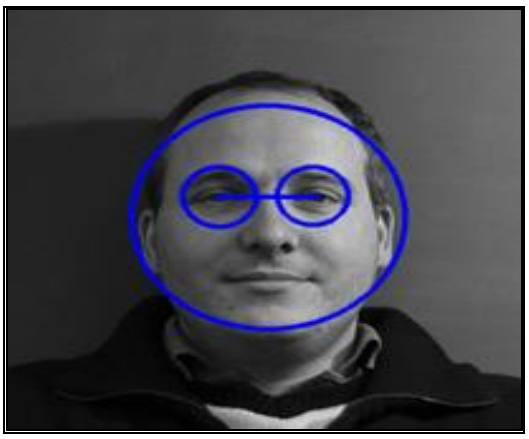
A etapa de detecção de faces e olhos é realizada com o uso do detector desenvolvido no trabalho de Markus et al. (2014), denominado *Pixel Intensity Comparisons Organized* – PICO. Trata-se de uma modificação no *framework* de detecção de objetos de Viola e Jones (2001), com base em um conjunto otimizado de árvores de decisão organizadas em uma cascata de classificadores.

Adicionalmente, existem dois métodos desenvolvidos em 2015 (LI et al., 2015; FARFADE, SABERIAN e LI-JIA, 2015), os quais obtiveram

resultados superiores ao método PICO na tarefa de detecção de faces. No entanto, os autores não disponibilizaram o código-fonte. Outro ponto é que apenas Li et al. (2015) disponibilizaram um executável em ambiente *Windows* para a execução de experimentos comparativos. Estes fatos, inviabilizaram o treinamento e a otimização dos referidos métodos. Por outro lado, uma implementação do método PICO foi retreinada com mais variações de pose de faces, o que produziu um detector mais preciso que o CascadeCNN (LI et al., 2015), conforme apresentado no próximo capítulo.

A correção da orientação das faces é feita pela rotação da imagem em torno do ponto central. O angulo de rotação é aquele formado pela reta que une o centro dos olhos e o eixo horizontal. O processo utilizado para detectar os olhos assumiu que a região de busca era composta de uma face previamente dectada. O detector de olhos fornecido na biblioteca OpenCV foi treinado com imagens de tamanho 18x12 pixels e seguiu a mesma estratégia de janela deslizante do detector de faces PICO (MARKUS et al., 2014). Uma ilustração de aplicação deste processo é apresentada na Figura 3.4.

Figura 3.4 – Etapa de detecção e correção da orientação de faces.



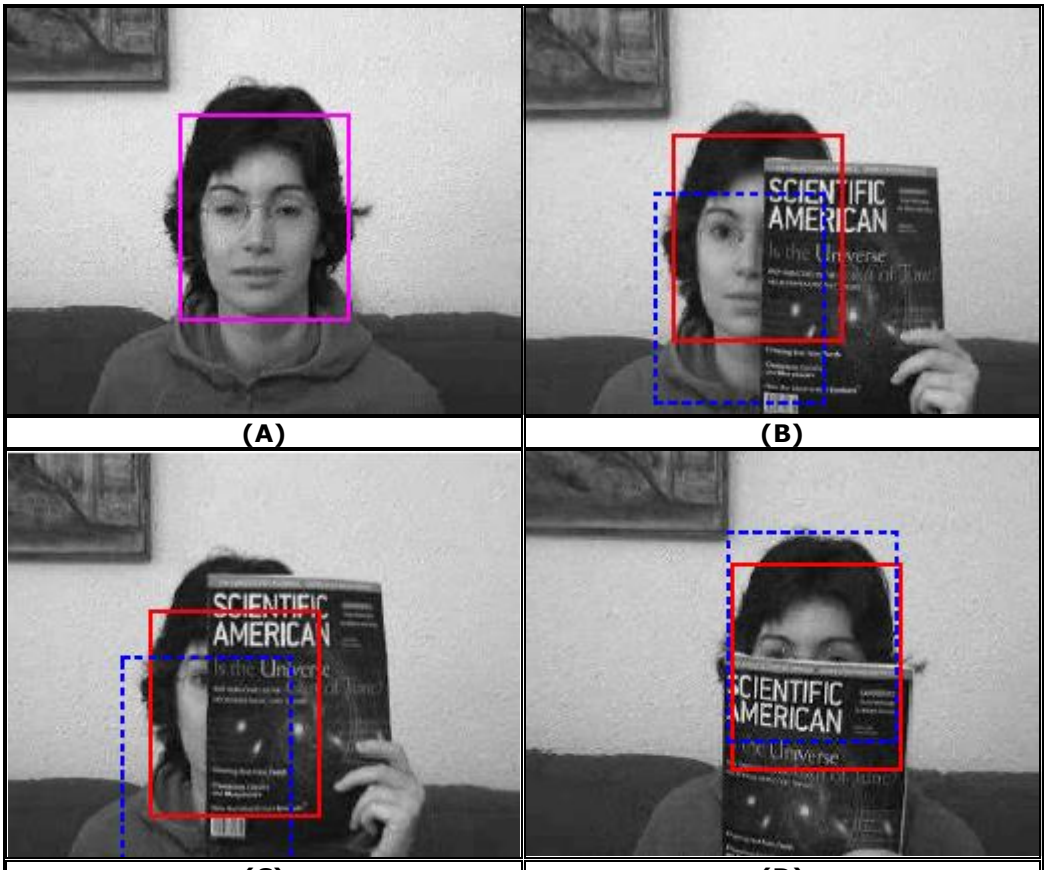
3.3.2. Rastreamento de Faces

A etapa de rastreamento de faces tem como objetivo auxiliar a etapa de detecção de faces, visando reduzir o o tempo de processamento da mesma. No rastreamento é realizada uma varredura por toda a imagem e em diferentes resoluções. A etapa de detecção de faces é executada a partir dos *shots* previamente identificados pelo processo descrito na Seção 3.2.1.

Uma vez que uma face é detectada no quadro que representa o *shot* – detectado pela ferramenta de Apostolidis e Mezaris (2014), o rastreador de faces, denominado *Frag*, desenvolvido por Adam, Rivlin e Shimshoni (2006) é executado a partir do quadro em que foi detectada a face até o final da cena, originando o *face tracklet*. Quando a cena é finalizada por um novo *shot*, a etapa de deteção e rastreamento é reiniciada. O processo é finalizado quando todas as cenas são processadas.

Na Figura 3.5, é ilustrado um exemplo comparativo do rastreador *Frag* (vermelho) com o rastreador *Tracking-Learning-Detection* – TLD (azul tracejado) (KALAL, MIKOLAJCZYK e MATAS, 2012). Para mais informações acerca do rastreador TLD vide Seção C.6.

Figura 3.5 – (A) Template inicial (face detectada); (B) Quadro #222; (C) Quadro #539; e (D) Quadro #849.



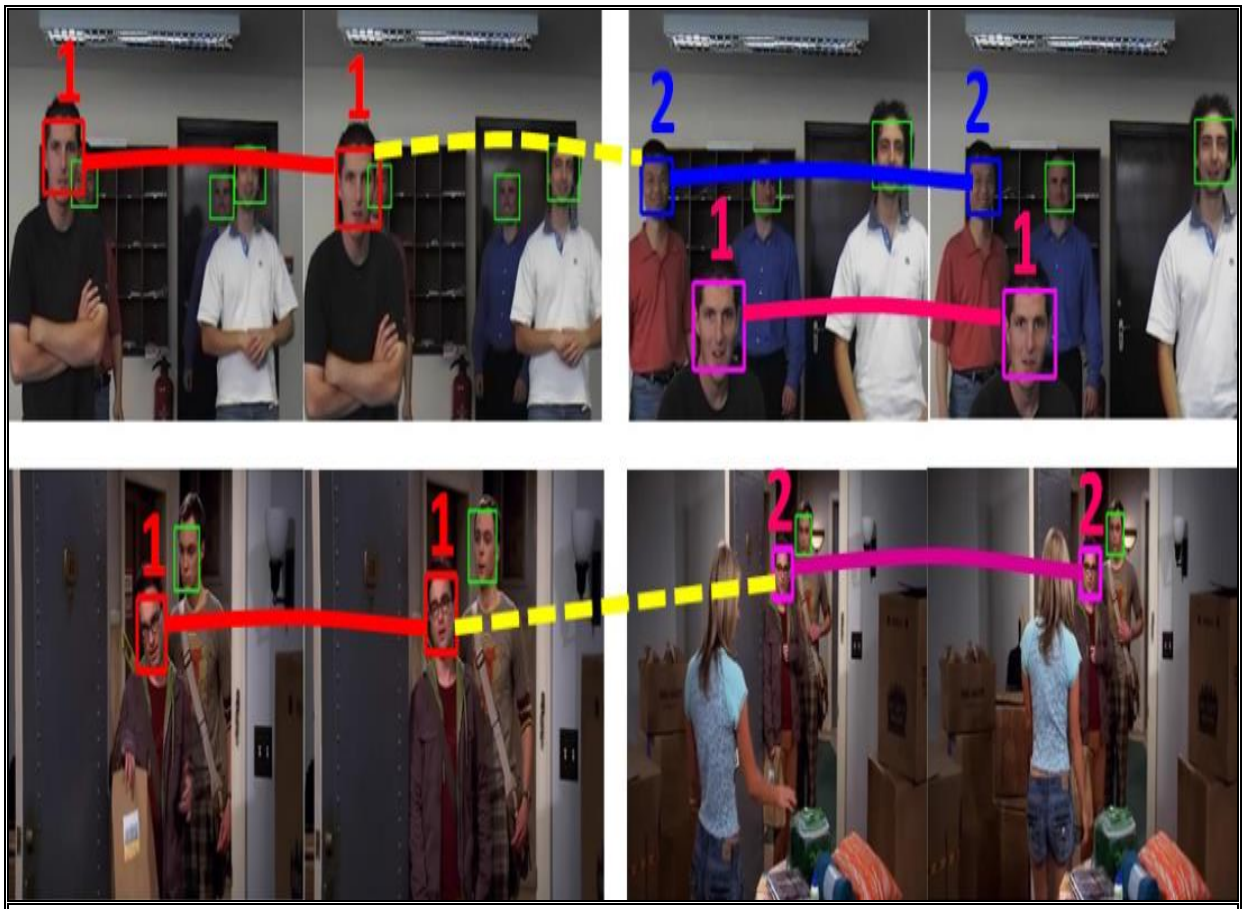
Fonte: Adaptado de Adam, Rivlin e Shimshoni (2006).

A partir da sequência de localizações das faces determinadas pelos rastreadores, são originados os *face tracklets*. Com base nas informações espaciais fornecidas pelos *face tracklets,* uma primeira etapa de agrupamento é realizada, com o intuito de vincular rastros da mesma

pessoa com diferentes variações de pose, iluminação ou oclusão. Nesta etapa, diferentes rastros da mesma pessoa são instanciados, melhorando o desempenho dos rastreadores de faces.

Na Figura 3.6, exemplificam-se os benefícios da colaboração entre agrupamento e rastreamento de faces: no primeiro caso (linha superior), vincular *tracklets* sem considerar o resultado do agrupamento leva à incorreta associação entre os rastros *#1* e *#2* ao considerar apenas a informação da localização das faces; e, no segundo caso (linha inferior), a semelhança entre as faces, retornada pelo agrupamento leva à correta associação entre os rastreadores *#1* e *#2*.

Figura 3.6 – Benefícios da colaboração entre agrupamento e rastreamento de faces.



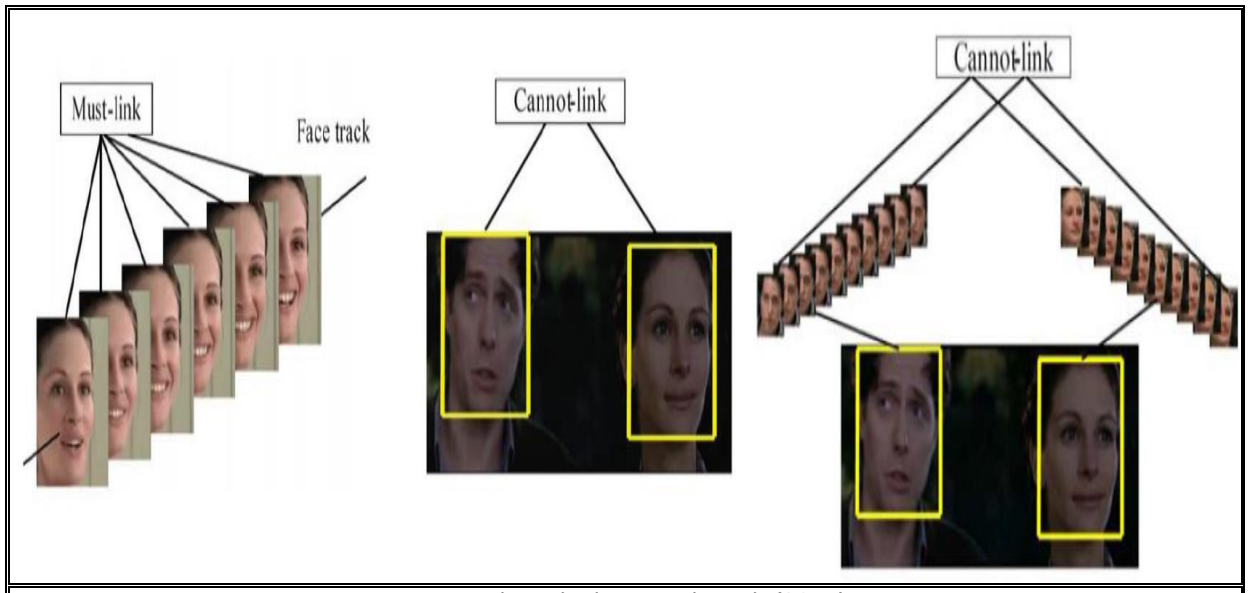
Fonte: Adaptado de Wu et al. (2013).

Com base nos grupos formados após o agrupamento dos *face tracklets*, são selecionadas amostras de faces representativas de cada grupo para a extração de características faciais, visando à formação de representações faciais por meio do uso de descritores SURF que servirão de entrada para a próxima etapa. Tal seleção faz-se necessária devido à

grande variabilidade de amostras de baixa qualidade (ocasionadas por mudanças de iluminação, pose, oclusão, etc.) que tendem a heterogeneizar o conjunto, diminuindo sua similaridade intergrupo e afetando negativamente o processo de agrupamento.

Sequências de vídeo contendo faces representativas podem fornecer informações adicionais para o processo de agrupamento. Em primeiro lugar, faces que aparecem simultaneamente no mesmo quadro não podem pertencer à mesma pessoa, constituindo os chamados *cannot-links*. Em segundo lugar, as faces selecionadas de um mesmo *face tracklet* devem pertencer à mesma pessoa, formando os chamados *must-links* (ver Figura 3.7). Esses dois tipos de restrições temporais são utilizados em conjunto com a matriz de similaridade para controlar o processo de agrupamento de um método não-supervisionado, como por exemplo, o método *Hierarchical Agglomerative Clustering* (HAC).

Figura 3.7 – Exemplos de restrições *cannot-link* e *must-link*.



Fonte: Adaptado de Harandi et al. (2011).

Apesar da vinculação dos *face tracklets* realizada pela etapa de agrupamento, variações substanciais de uma mesma face podem ocorrer e consequentemente, irão gerar uma cisão no agrupamento de faces de uma mesma pessoa. A etapa final da abordagem proposta é responsável por realizar um segundo processo de agrupamento, com base em uma ordenação de similaridade entre os *face tracklets* e uma fusão dos grupos de faces similares que possam ter sofrido cisão.

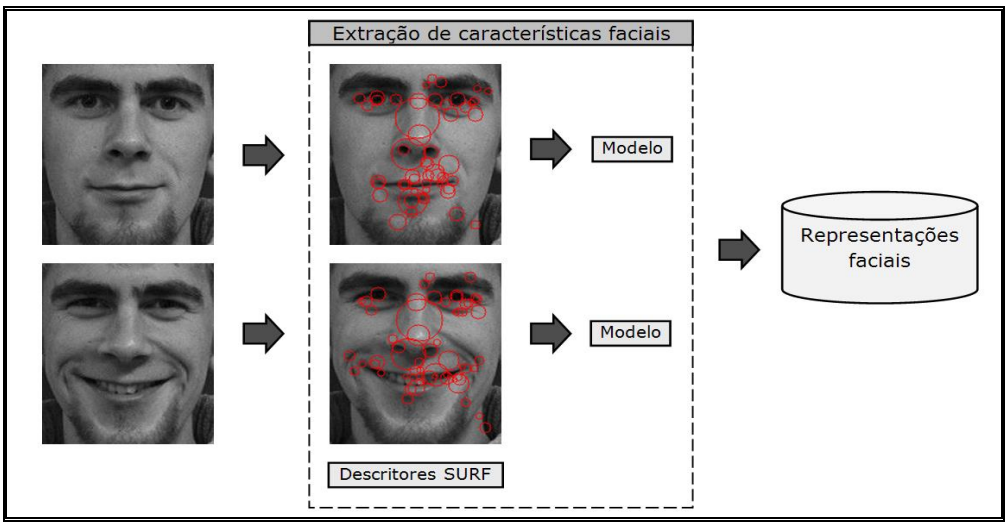
3.3.3. Extração de Características Faciais

Ao se tentar comparar características entre imagens diferentes, frequentemente se é confrontado com o problema das mudanças de escala, ou seja, as diferentes imagens a serem analisadas podem ser tomadas a diferentes distâncias dos objetos de interesse e, consequentemente, esses objetos serão retratados em diferentes tamanhos. Ao se tentar extrair uma mesma característica de duas imagens usando uma vizinhança de tamanho fixo, por causa da mudança de escala, os seus padrões de intensidade não corresponderão (LAGANIÈRE, 2011).

Para resolver esse problema, o conceito de características invariantes à escala foi introduzido na área de Visão Computacional (VC). Nos últimos anos, duas abordagens promissoras para detectar regiões salientes em imagens foram desenvolvidas: *Scale Invariant Feature Transform* – SIFT (LOWE, 1999) e *Speeded Up Robust Features* – SURF (BAY, TUYTELAARS e VAN GOOL, 2006).

Ambas as abordagens não apenas detectam pontos de interesse (*keypoints*), como também propõem um método para criação de um descritor invariante. Esse descritor pode ser usado para identificar pontos de interesse únicos e compará-los, mesmo sob uma variedade de condições que dificultam o reconhecimento, tais como mudanças de escala, rotação, iluminação e pose (BAUER, SÜNDERHAUF e PROTZEL, 2007). Uma ilustração de aplicação desta técnica é apresentada na Figura 3.8.

Figura 3.8 – Exemplo de extração de características faciais SURF.

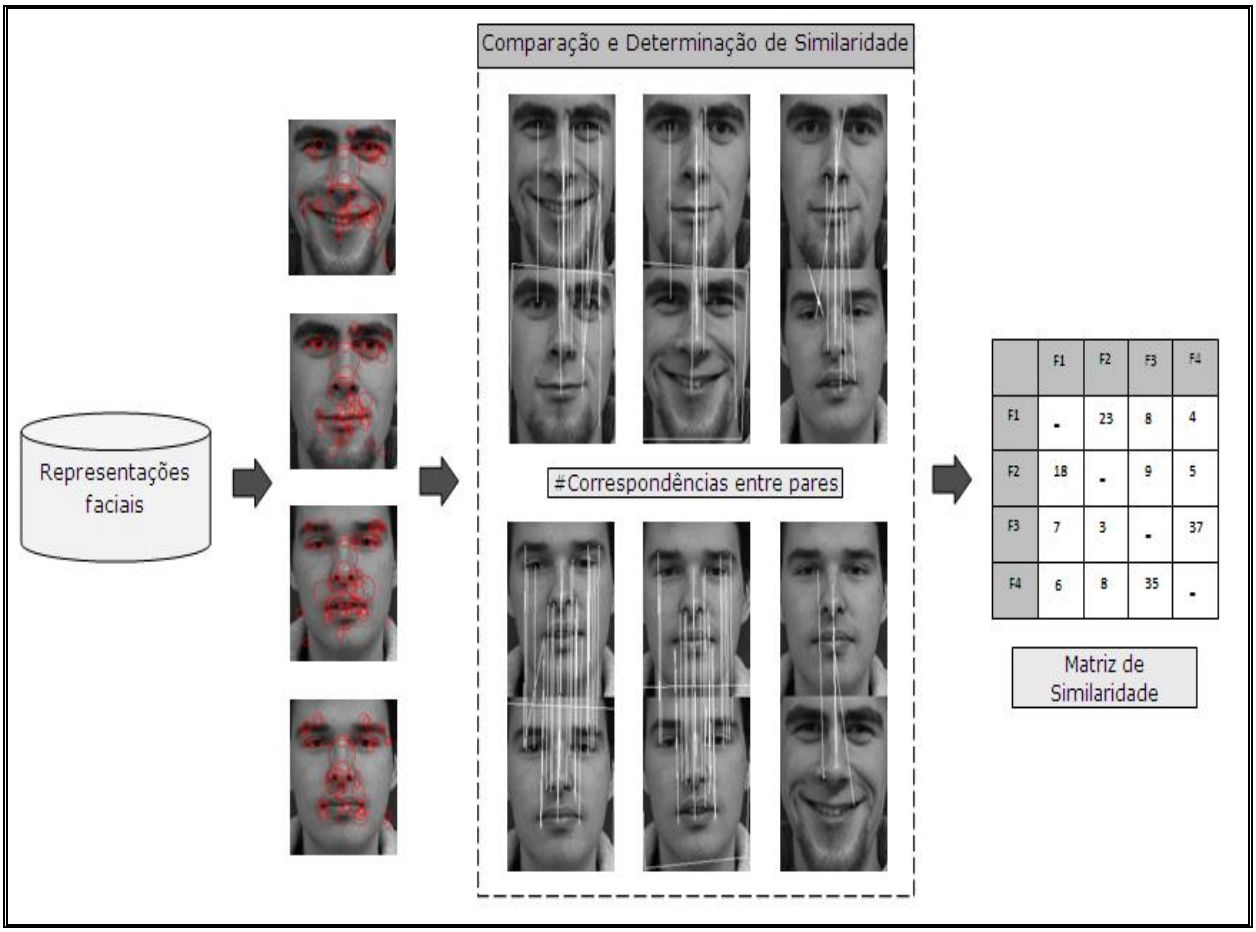


Conforme mencionado na Seção 2.5, experimentos comparativos realizados por Juan e Gwon (2009) demonstraram vantagens ao se utilizar o SURF em comparação com o SIFT na base de imagens Caltech (CALTECH FACE DATABASE, 2010). A partir desse resultado, o descritor SURF foi escolhido como a técnica para extração de características faciais utilizada neste trabalho.

3.3.4. Comparação e Determinação de Similaridade

De posse das representações faciais extraídas por meio dos descritores SURF das faces representativas, o módulo de *comparação e determinação de similaridade* é responsável pela determinação da semelhança entre os elementos do conjunto de todos possíveis pares de faces a serem agrupadas. Na Figura 3.9, ilustra-se o detalhamento funcional deste módulo.

Figura 3.9 – Comparação e determinação de similaridade das representações faciais.



O módulo de *comparação e determinação de similaridade* utiliza o algoritmo proposto por Muja e Loew (2009), denominado *Fast Approximate*

Nearest Neighbors – FANN (vide Apêndice C), para determinar o grau de similaridade entre duas faces em função da correspondência entre descritores e inspirado na abordagem proposta por Antonopoulos, Nikolaidis e Pitas (2007), a qual é utilizada para mapear esta correspondência em uma matriz de similaridade.

O algoritmo FANN (vide a Seção 3 do Apêndice C) tem como saída o número de correspondências entre os pontos de interesse (*keypoints*), calculados pelo descritor SURF do par de descritores comparados. A partir destas correspondências, histogramas de cinza são extraídos dos pontos de interesse que compõem cada correspondência e ponderados por sua interseção, conforme a métrica proposta por Swain e Ballard (1991):

$$\boldsymbol{d(H_p, H_q) = \frac{\sum_i min(H_p^i, H_q^i)}{\sum_i H_p^i}} \qquad \textbf{(3.1)}$$

na qual H_p e H_q são os histogramas de cinza dos pontos de interesse (*keypoints*) de cada uma das correspondências. Essa função assume um valor real que varia de 0, para situações em que os histogramas são totalmente diferentes; até 1, quando idênticos.

Devido ao fato de que a correspondência entre os descritores A e B não produz o mesmo resultado da correspondência entre B e A e considerando que a matriz de similaridade deve ser simétrica, faz-se necessário o cálculo da correspondência duas vezes para cada par de descritores, ou seja, uma vez para o par (A, B) e outra vez para o par (B, A).

Assim, o número máximo de correspondências encontradas entre (A, B) e (B, A) é considerado como a correspondência final para este par de descritores. Desta forma, a matriz de similaridade pode ser definida de acordo com a seguinte função de similaridade proposta por Antonopoulos, Nikolaidis e Pitas (2007):

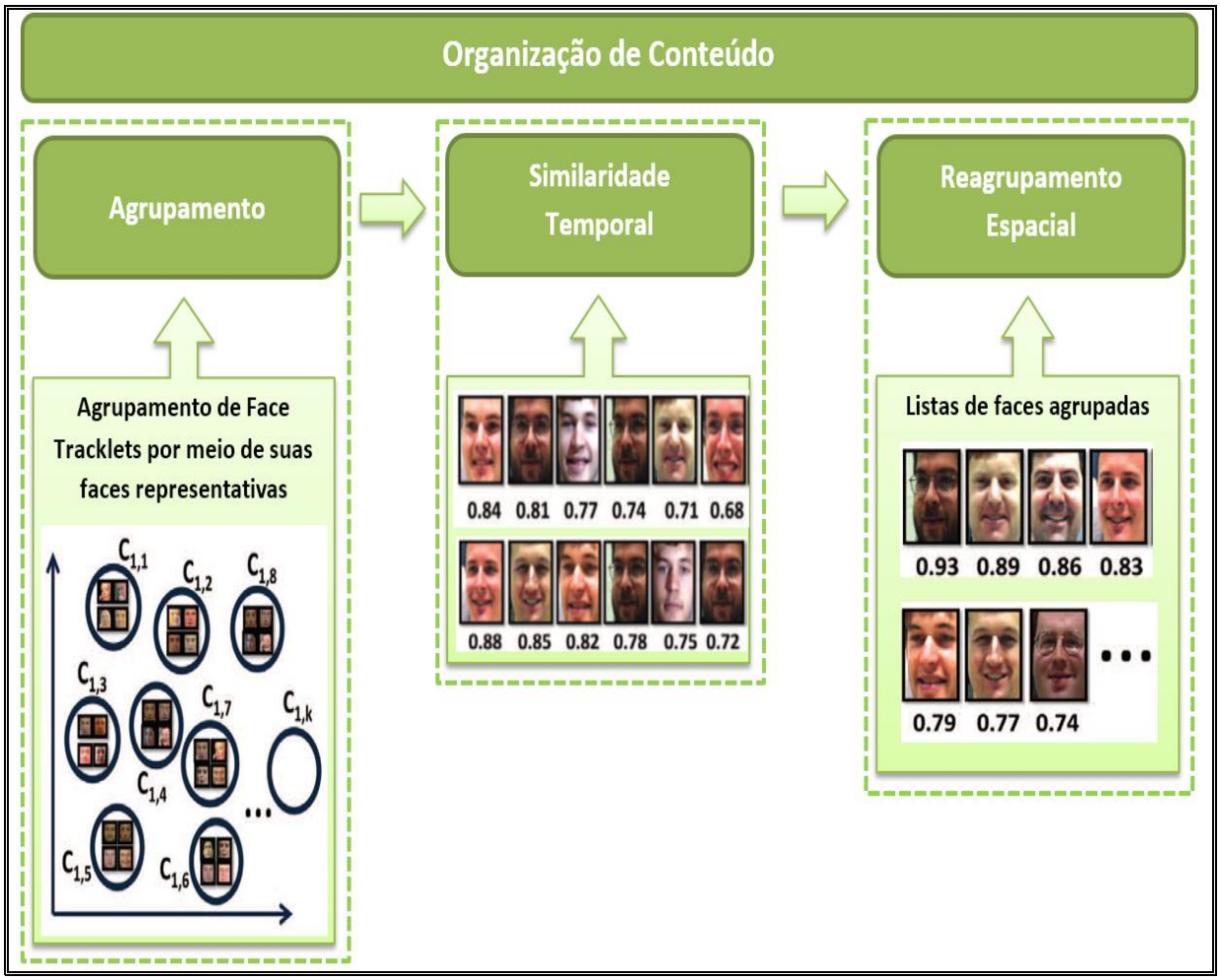
$$\boldsymbol{S(A,B) = S(B,A) = 100\left(1 - \frac{M_{AB}}{\min(K_A, K_B)}\right)}, \qquad \textbf{(3.2)}$$

em que M_{AB} é o número máximo de correspondências entre (A, B) e (B, A) e K_A e K_B são o número de pontos de interesse dos descritores A e B, respectivamente. Essa função assume valores no intervalo [0,100], em que valores próximos de 0 indicam grande similaridade entre os descritores. A matriz de similaridade computada servirá de entrada para a próxima etapa de processamento do sistema proposto, o módulo de organização de conteúdo.

3.4. Organização de Conteúdo

O módulo final do sistema trata do cerne de processamento onde as listas de faces agrupadas de saída são geradas. Esta etapa da abordagem proposta compreende três etapas, a saber: (i) agrupamento de *face tracklets*; (ii) similaridade temporal; e (iii) reagrupamento espacial, conforme ilustrado na Figura 3.10.

Figura 3.10 – Visão detalhada do módulo de organização de conteúdo.



A partir das faces representativas extraídas dos *face tracklets*, a etapa

final do sistema proposto, denominada *Organização de Conteúdo*, objetiva a composição dos grupos de faces semelhantes utilizando um agrupamento aglomerativo hierárquico espaço-temporal. Deste modo, os pares de faces são arranjados de acordo com os respectivos graus de similaridade e informações espaço-temporais, agrupados segundo uma disposição hierárquica.

Os métodos hierárquicos (vide Seção 9 do Apêndice C) são métodos simples em que os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (EVERITT, LANDAU e MORVEN, 2001). Essa representação facilita visualizar a formação dos agrupamentos em cada estágio no qual ela ocorreu, assim como o grau de semelhança entre eles.

Diante do exposto, a etapa final da abordagem proposta utiliza incialmente o método HAC sobre as faces representativas extraídas dos *face tracklets*. Em seguida, com base na informação temporal os grupos gerados são ordenados de maneira ascendente de acordo com um limiar t, que representa a diferença de tempo máxima entre duas faces representativas para que as mesmas pertençam ao mesmo grupo.

Assim, a determinação da similaridade temporal dos grupos de faces representativas pode ser formalizada como:

(1) Definir a diferença de tempo máxima t;

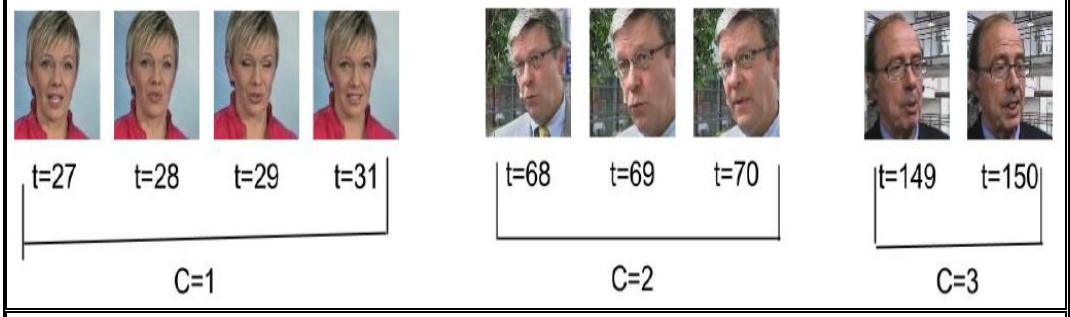
(2) Ordenar de maneira ascendente pela ordem cronológica os grupos de *face tracklets*;

(3) Inicializar o índice dos novos grupos como $C = 1$;

(4) Atribuir ao primeiro grupo ($C = 1$) as faces representativas ao primeiro espaço de tempo; e

(5) Verificar a diferença entre os espaços de tempo consecutivos. Caso a diferença seja menor do que t, adicionar ao grupo corrente (C) as faces representativas consecutivas. Caso contrário, incrementar o índice de C criando um novo grupo e associar as faces representativas em questão (ver Figura 3.11).

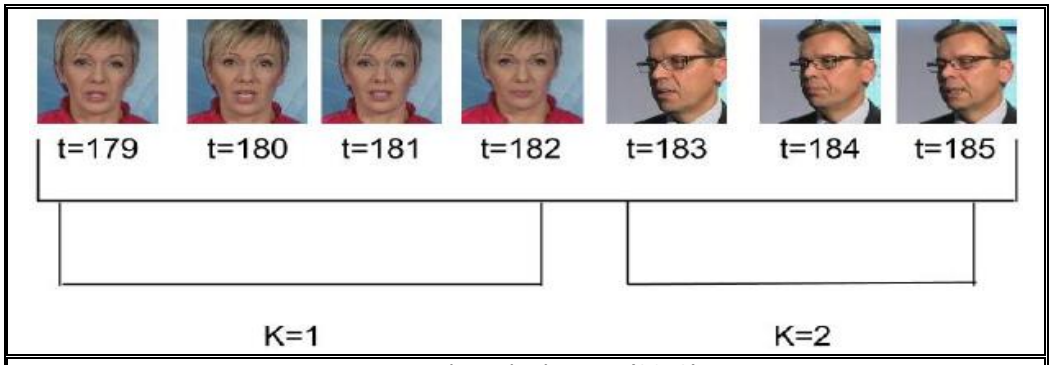
Figura 3.11 – Similaridade temporal de grupos de faces representativas.



Fonte: Adaptado de Mian (2013).

Após a etapa de similaridade temporal, os grupos formados passam por uma etapa de reagrupamento com base na informação espacial das faces representativas obtidas dos *face tracklets*. Neste passo, considera-se o fato de que cada quadro do vídeo pode ter a presença de duas ou mais faces, em diferentes coordenadas espaciais. Assim, cada grupo temporal poderá conter faces representativas de duas ou mais pessoas, cujas coordenadas espaciais são diferentes. Esta informação é utilizada, a fim de melhorar ainda mais o mecanismo de inicialização para uma segunda execução do algoritmo HAC, conforme ilustrado na Figura 3.12.

Figura 3.12 – Reagrupamento com base na informação espacial.



Fonte: Adaptado de Mian (2013).

3.5. Considerações Finais

Neste capítulo, foi proposta uma abordagem para o agrupamento de faces em vídeos digitais, a qual é dividida em três módulos: (i) preparação de conteúdo; (ii) seleção de conteúdo; e (iii) organização de conteúdo.

Na etapa inicial do sistema proposto nesta tese, os quadros do vídeo de entrada são extraídos e submetidos a um pré-processamento de correção de iluminação (filtragem homomórfica) e equalização de histograma. Após o

processo de pré-processamento, as etapas de detecção/rastreamento de faces são realizadas a fim de identificar as faces presentes nos quadros para posterior extração de características. Descritores SURF (BAY, TUYTELAARS e VAN GOOL, 2006) são utilizados para transformar as imagens de face em representações faciais na forma de pontos de interesse, assim como a composição dos *face tracklets* e faces representativas que servirão de entrada para o próximo módulo do sistema. Ocorre uma comparação de similaridade entre todas as faces representativas, a qual permite gerar uma matriz de similaridade e uma primeira etapa de agrupamento é realizada. Por fim, as informações espaço-temporal são utilizadas para refinar o processo de agrupamento dando origem a listas de faces agrupadas, como saída do sistema.

O próximo capítulo contém uma descrição dos experimentos realizados neste trabalho e discussão dos resultados obtidos.

Capítulo 4

Avaliação Experimental

Neste capítulo são descritos e comentados os experimentos realizados para a avaliação da abordagem proposta para a solução do problema de agrupamento de faces em vídeos, apresentada no Capítulo 3. Os componentes da abordagem proposta foram analisados e comparados, por meio de testes numéricos e visuais, com outros resultados provenientes de abordagens concorrentes.

Os experimentos foram conduzidos em etapas distintas do estudo, sendo cada uma dessas etapas descrita em seções separadas, a saber:

- Na Seção 4.1, são apresentados resultados de comparação de três abordagens para detecção de faces: *CascadeCNN* (LI et al., 2015), *PICO* (MARKUS et al., 2014) e *PICO TRAIN* (versão do detector *PICO* com treinamento aperfeiçoado nesta Tese para acomodar faces de perfil). Os experimentos foram realizados nas bases de dados *Face Detection Data Set and Benchmark* – FDDB (JAIN e LEARNED-MILLER, 2010) e *YouTube Celebrities* (KIM et al., 2008);
- Na Seção 4.2, são apresentados resultados da etapa de rastreamento de faces. O experimento realizado por Wang et al. (2014) foi estendido para avaliação do rastreador SURF proposto nesta Tese. No experimento foram consideradas três sequências de vídeo que apresentam variações de pose, iluminação e oclusão;
- Na Seção 4.3, são apresentados resultados comparativos para a tarefa de agrupamento de faces em vídeos. A partir de métricas

de avaliação de agrupamento, inicialmente, o método proposto foi avaliado com base na combinação de componentes internos, a fim de mensurar a contribuição de cada componente no resultado final do sistema. Em seguida, a melhor combinação de componentes do método proposto foi comparada com alguns trabalhos concorrentes nas bases de dados *YouTube Celebrities* (KIM et al., 2008) e *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013); e, por fim,

- Na Seção 4.4, os resultados obtidos são analisados e discutidos sob a forma de considerações finais.

4.1. Detecção de Faces

A detecção de faces é utilizada para identificar a localização e o tamanho de cada face presente em uma imagem de entrada. Quando empregada em conjunto com um rastreador de faces, a detecção de faces atua na etapa de inicialização do rastreador, podendo ser novamente executada a partir de uma quantidade pré-determinada de quadros, a fim de corrigir o rastro da face que está sendo rastreada.

Os detectores comparados nos experimentos foram *CascadeCNN* (LI et al., 2015), *PICO* (MARKUS et al., 2014) e *PICO TRAIN* (versão do detector *PICO* com treinamento customizado na presente pesquisa). Tais detectores foram selecionados por apresentarem código fonte disponível e implementação na linguagem C/C++ possibilitando uma integração direta com o sistema implementado nesta Tese.

Adicionalmente, apesar de uma grande quantidade de trabalhos serem relacionados com o problema de detecção de faces, optou-se pela utilização dos detectores *CascadeCNN* (LI et al., 2015) e *PICO* (MARKUS et al., 2014) devido os seguintes fatores:

(1) **Necessidade de implementação**: dado que existem inúmeros detectores que poderiam ser empregados na avaliação, a implementação de uma certa quantidade dos trabalhos iria demandar um tempo considerável, pois uma pequena parcela dos

autores fornece a implementação (binário) de sua abordagem, o que poderia afetar o cronograma planejado;

(2) **Necessidade de validação**: a partir da uma hipotética implementação dos trabalhos relacionados, testes de validação seriam necessários para verificar a fidelidade do código, a fim de evitar que resultados incorretos pudessem ser utilizados erroneamente;

(3) **Critérios de avaliação diferentes**: partindo do pressuposto de que os fatores 1 e 2 não pudessem ser realizados, uma alternativa seria a utilização de resultados reportados nos trabalhos publicados; no entanto, devido ao fato de que a grande maioria dos trabalhos analisados adotou critérios (supervisionado ou não-supervisionado) e métricas de avaliação diferentes, isto impossibilitou uma comparação direta com os detectores escolhidos; e

(4) **Utilização prévia**: os dois detectores de faces consideradas no experimento também foram utilizados com sucesso em experimentos preliminares, o que favoreceu seu uso, por conta da familiarização com suas funcionalidades.

4.1.1. Bases de Dados

Na avaliação experimental do detector de faces adotado na composição do método proposto nesta Tese, foram utilizadas as bases de imagens *Face Detection Data Set and Benchmark* – FDDB (JAIN e LEARNED-MILLER, 2010) e *YouTube Celebrities* (KIM et al., 2008).

A FDDB (JAIN e LEARNED-MILLER, 2010) é uma base de referência na área, além de apresentar um processo de avaliação padronizado. No escore de avaliação discreto é calculado o número de faces detectadas sobre o número de falsas detecções. Uma detecção é considerada como "verdadeiro positivo" apenas se razão *S* entre a região detectada e o *ground-truth* for igual ou superior a 0,5 (JAIN e LEARNED-MILLER, 2010), conforme a Equação (4.1).

$$S(g_i, t_i) = \frac{area(g_i) \cap area(t_i)}{area(g_i) \cup area(t_i)} \quad \textbf{(4.1)}$$

em que g_i corresponde à região do *ground-truth* da *i*-ésima imagem e t_i corresponde à região detectada da *i*-ésima imagem.

A base de imagens FDDB contém 5,171 faces anotadas (em forma de elipse) em 2,845 imagens (vide Figura 4.1 para alguns exemplos).

Figura 4.1 – Amostras de imagens da base FDDB.

Fonte: Adaptado de Jain e Learned-Miller (2010).

A base de imagens *YouTube Celebrities* (KIM et al., 2008) contém 1,910 sequências de vídeo de 47 celebridades (atores, esportistas e políticos). Cada sequência possui centenas de quadros de baixa resolução e alta taxa de compressão (codificados em MPEG4 a uma taxa de 25 quadros por segundo), conforme ilustrado na Figura 4.2. O critério de avaliação definido para a base FDDB (JAIN e LEARNED-MILLER, 2010) foi adotado para a base *YouTube Celebrities* (KIM et al., 2008).

Figura 4.2 – Amostras de imagens da base *YouTube Celebrities*.

Fonte: Adaptado de Kim et al. (2008).

4.1.2. Resultados Experimentais

O detector *PICO Train* foi treinado a partir de 13,233 imagens frontais e semi-perfil da base LFW (HUANG et al., 2007) e 4,797 imagens de perfil e semi-perfil das bases UMIST (GRAHAM e ALLINSON, 1998), CMU Mobo (GROSS e SHI, 2001), *Oulu Face Video Database* (MARTINKAUPPI et al., 2002) e CMU PIE (SIM, BAKER e BSAT, 2003), totalizando 18,030 imagens de faces. Adicionalmente, para o conjunto de exemplos negativos, que não contém faces, foram utilizadas 300,000 imagens[5] fornecidas pelos autores da técnica PICO (MARKUS et al., 2014).

Para determinar qual o melhor detector e, consequentemente, as melhores taxas de detecção foram geradas curvas ROC - *Receiver Operating Characteristic* (BRADLEY, 1997). Para a avaliação objetiva do desempenho de detecção, foram calculados os valores da área sob a curva (AUC) e o erro

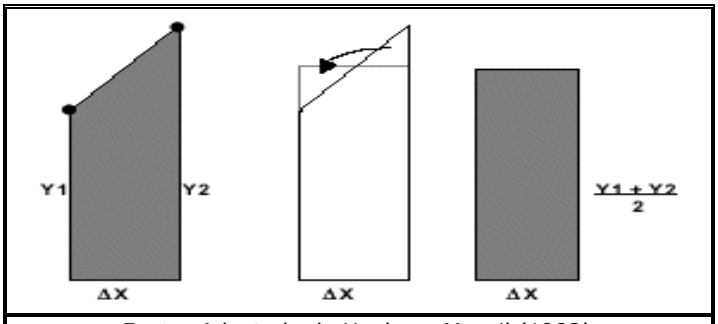
[5] Imagens disponíveis em: *https://github.com/nenadmarkus/pico*

padrão (SE) (HANLEY e MCNEIL, 1982).

Vale ressaltar que no cenário de detecção de faces o eixo das abscissas (X) normalmente é absoluto, pois a taxa de FPR é muito pequena, tipicamente 1/1.000.000. Assim, considera-se o valor do somatório de falsos positivos.

O valor da AUC é calculado pelo método do trapézio composto. A área sob cada um dos segmentos conectados descreve um trapézio, conforme ilustrado na Figura 4.3. Assim, a área de cada trapézio é calculada pela da área do retângulo equivalente, sendo a AUC o somatório das áreas de todos os retângulos.

Figura 4.3 – Área de um trapézio calculada pela área do retângulo equivalente.



Fonte: Adaptado de Hanley e Mcneil (1982).

Desta forma, o erro padrão (SE) da AUC é calculado pela Equação (4.2) (HANLEY e MCNEIL, 1982):

$$SE = \sqrt{\frac{A(1-A) + (np-1)(Q1 - A*A) + (nn-1)(Q2 - A*A)}{np*nn}} \qquad \textbf{(4.2)}$$

em que *A* corresponde ao valor da AUC, *np* e *nn* correspondem aos números de amostras positivas e negativas, respectivamente, *Q1* e *Q2* são calculados por:

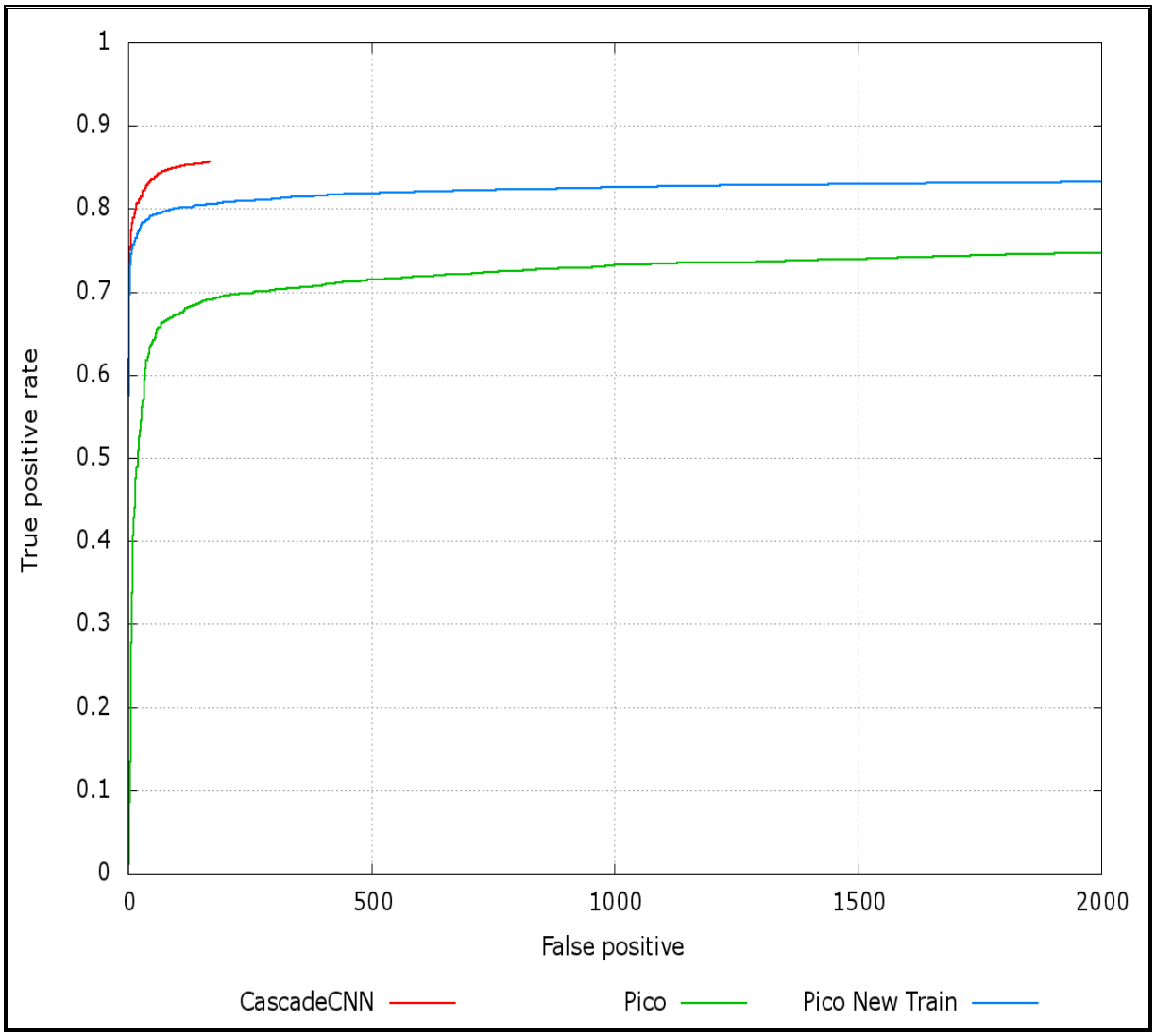
$$Q1 = \frac{A}{2-A}, \quad Q2 = \frac{2A*A}{1+A} \qquad \textbf{(4.3)}$$

Na Figura 4.4, é ilustrado o desempenho dos detectores avaliados na base de imagens FDDB (JAIN e LEARNED-MILLER, 2010) que consiste de 5171 faces presentes em 2845 imagens retiradas da base LFW (HUANG et al., 2007), sendo o desempenho do detector *CascadeCNN* (LI et al., 2015) superior aos demais. Todas as 5171 imagens foram utilizadas na avaliação

dos classificadores.

Observando-se a Figura 4.4, percebe-se que a curva ROC do detector *CascadeCNN* (LI et al., 2015) não possui valores de falso positivo superiores a 160, indicando que, possivelmente, nem todos os parâmetros do detector foram exaustivamente investigados.

Figura 4.4 – Curvas ROC dos três detectores avaliados na base FDDB.



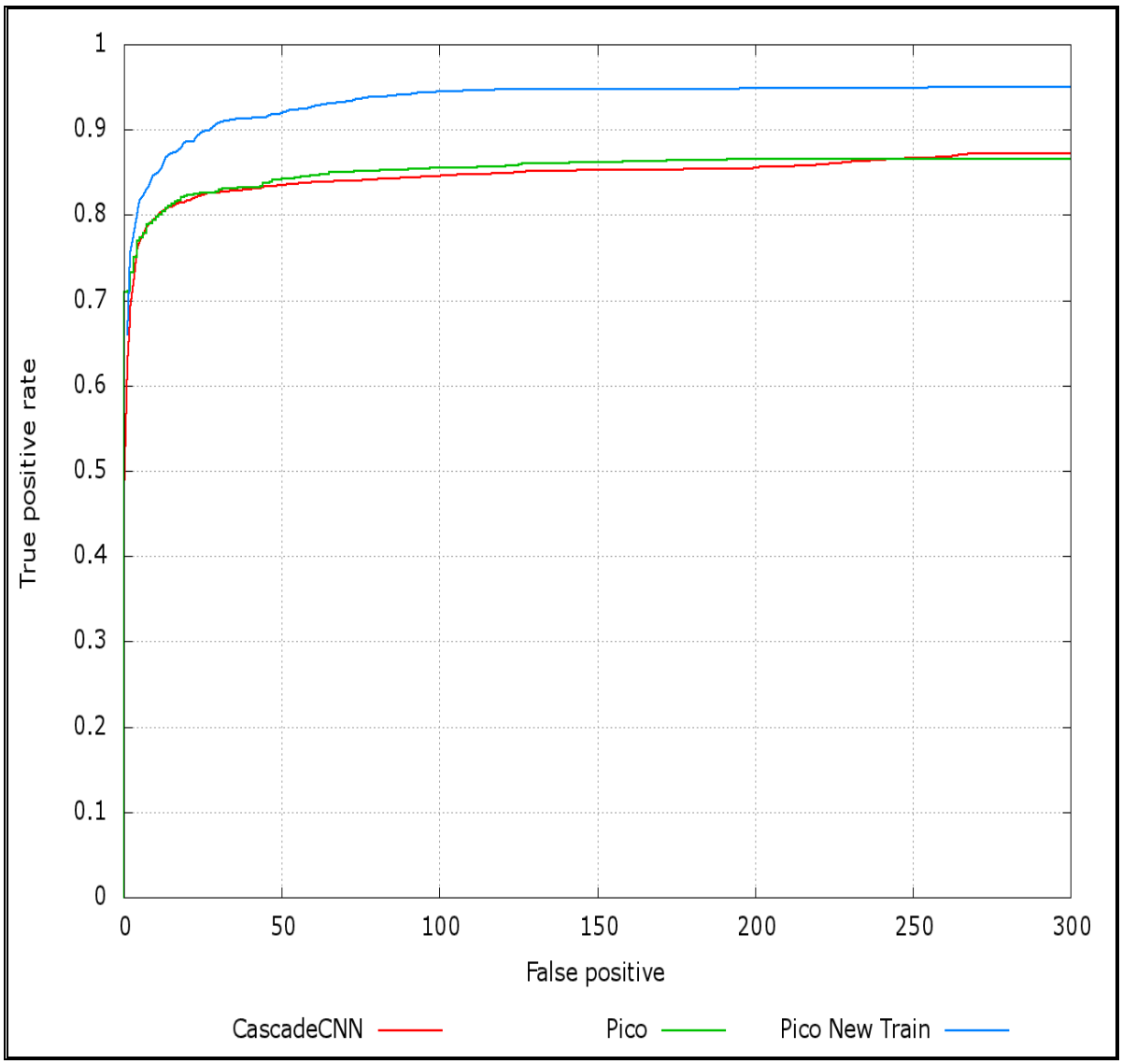
Adicionalmente, foram calculados os valores de AUC e SE para cada um dos três detectores avaliados na base de imagens FDDB (JAIN e LEARNED-MILLER, 2010), conforme ilustrado na Tabela 4.1.

Tabela 4.1 – Valores de AUC e SE de cada detector na base FDDB.

Detector	AUC	SE
CascadeCNN	**0,83566**	± 0,01132
PICO	0,70646	± 0,00608
PICO TRAIN	0,80941	± 0,00481

Na Figura 4.5, é ilustrado o desempenho dos detectores avaliados na base de vídeos *YouTube Celebrities* (KIM et al., 2008), neste experimento, todos os 1910 vídeos foram utilizados na avaliação, sendo o desempenho do detector PICO TRAIN superior aos demais.

Figura 4.5 – Curvas ROC dos três detectores avaliados na base *YouTube Celebrities*.



Adicionalmente, foram calculados os valores de AUC e SE para cada um dos três detectores avaliados na base de imagens *YouTube Celebrities* (KIM et al., 2008), conforme ilustrado na Tabela 4.2.

Tabela 4.2 – Valores de AUC e SE de cada detector na base *YouTube Celebrities*.

Detector	AUC	SE
CascadeCNN	0,87023	± 0,00385
PICO	0,85971	± 0,00804
PICO TRAIN	**0,94017**	± 0,00286

4.1.3. Conclusões

O detector PICO TRAIN foi selecionado para a integração ao sistema desenvolvido nesta tese, uma vez que apresentou um resultado com relação a AUC próximo (~2.5% inferior) ao detector CascadeCNN na base FDDB e o melhor resultado (~7.0% superior ao detector por CascadeCNN) na base *YouTube Celebrities*, sendo esta, uma das bases de referência para o problema de agrupamento de faces em vídeos, possuindo, assim, um peso maior na tomada de decisão.

4.2. Rastreamento de Faces

Nesta seção, são descritos e comentados os resultados obtidos no experimento de rastreamento de faces, um dos componentes da abordagem proposta para o problema de agrupamento de faces em vídeos. Na avaliação experimental foram considerados os resultados provenientes do trabalho de Wang et al. (2014), com adição do rastreador SURF (vide Apêndice C) proposto para efeito de comparação, em três sequências de vídeo definidas no estudo de Wang et al. (2014).

4.2.1. Base de Dados

Na avaliação experimental, foram utilizadas as sequências de vídeo *Face:bb*[1], *Face:mb*[6] e *Face:po*[7], apresentando uma variabilidade de pose, iluminação e oclusão. A sequência de vídeo *Face:bb* contém 50 quadros anotados onde estão presentes duas pessoas. A sequência de vídeo *Face:mb* contém 500 quadros anotados com duas pessoas presentes. Por fim, a sequência de vídeo *Face:po apresenta apenas uma pessoa em 898 quadros* (vide Figura 4.6 para alguns exemplos).

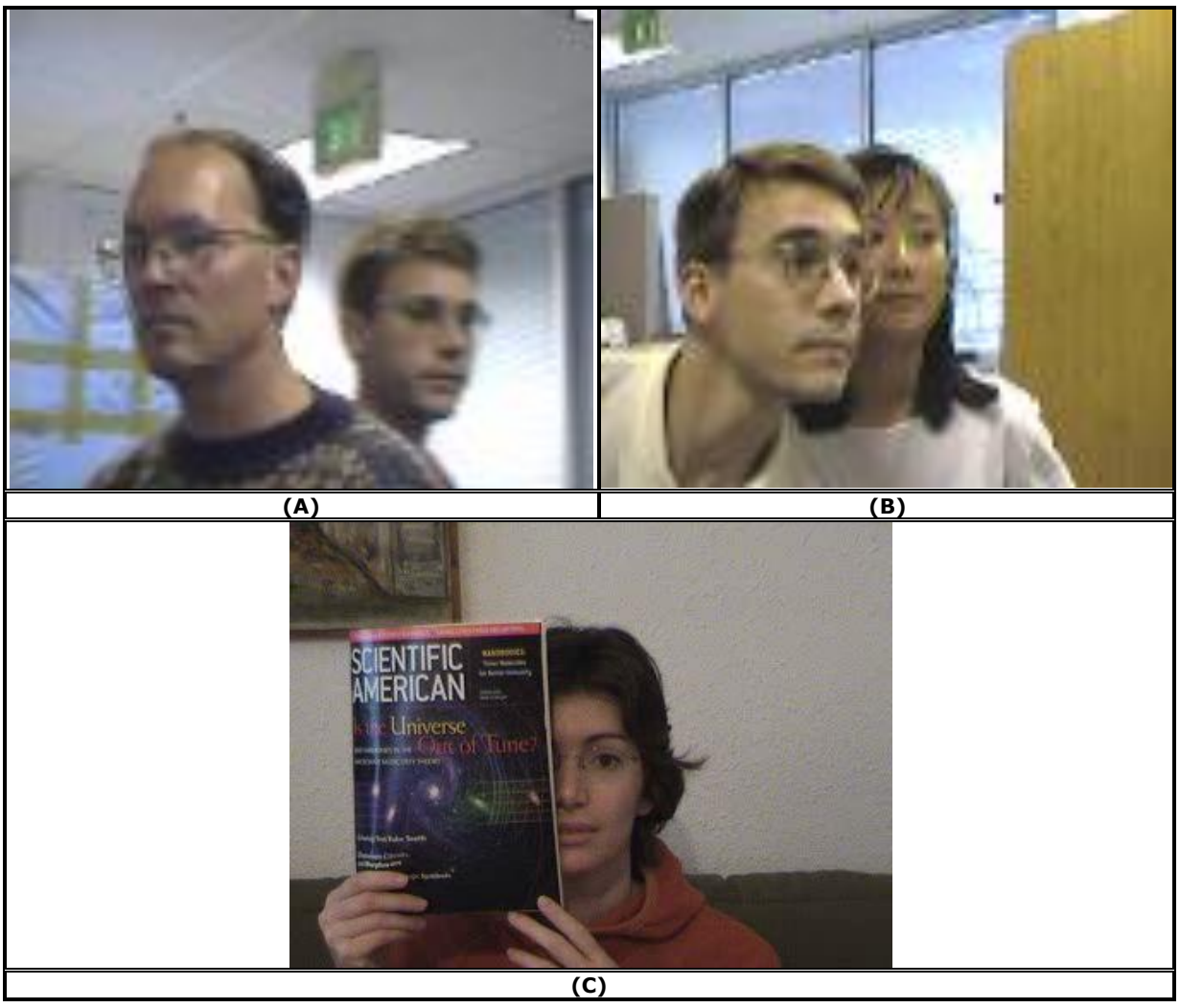
4.2.2. Resultados Experimentais

Os métodos comparados no experimento de rastreamento de faces foram TLD – *Tracking-Learning-Detection* (KALAL, MIKOLAJCZYK e MATAS, 2012), FRAG – *Robust fragments-based tracking using the integral histogram*

[6] Vídeos *Face:bb* e *Face:mb*, disponíveis em: *http://www.ces.clemson.edu/stb/research/headtracker/seq/*
[7] Vídeo *Face:po*, disponível em: *http://www.cs.technion.ac.il/~amita/fragtrack/fragtrack.htm*

(ADAM, RIVLIN e SHIMSHONI, 2006), SCM – *Robust object tracking via sparsity-based collaborative model* (ZHONG, LU e YANG, 2012) e SURF (vide Seção 4.2.1). Tais métodos foram escolhidos por apresentarem os melhores resultados nos experimentos relatados no trabalho de Wang et al. (2014).

Figura 4.6 – Amostras de quadros: (A) Face:bb; (B) Face:mb; e (C) Face:po.



(A) (B)

(C)

Para determinar qual o melhor rastreador, foram utilizadas as métricas *PosErr* (*position error*) e *F-score*. A métrica *PosErr* mede a diferença em pixels entre o centro da região da face do *ground-truth* e o centro da região rastreada, conforme a Equação (4.4) (WANG et al., 2014):

$$PosErr = \sqrt{\left(c_{trk}^{x} - c_{gt}^{x}\right)^2 + \left(c_{trk}^{y} - c_{gt}^{y}\right)^2} \quad \textbf{(4.4)}$$

em que ($c^x_{trk,}$ c^y_{trk}) e ($c^x_{gt,}$ c^y_{gt}) denotam a posição central rastreada e a posição central do ground-truth, respectivamente. A métrica *F-score* avalia a acurácia do rastreador levando em consideração a cobertura (*recall*) e

precisão (*precision*), conforme a Equação (4.5) (WANG et al., 2014):

$$F\text{–}score = \frac{TP * 2}{TP * 2 + FP + FN}, \tag{4.5}$$

em que *TP*, *FP* e *FN* denotam verdadeiro positivo, falso positivo e falso negativo, respectivamente. No cenário de rastreamento, a métrica *F-score* pode ser convertida em uma formulação equivalente (WANG et al., 2014), conforme especificado pela Equação (4.).

$$F\text{–}score = \frac{R_{trk} \cap R_{gt} * 2}{R_{trk} \cap R_{gt} + R_{trk} \cup R_{gt}}, \tag{4.6}$$

em que R_{trk} denota a região da face sendo rastreada e R_{gt} a região da face no *ground-truth*. Um bom resultado de rastreamento deve fornecer um baixo valor de *PosErr* e um alto valor de *F-score*.

Na Tabela 4.3, são apresentados os resultados dos quatro rastreadores para a métrica *PosErr,* para cada uma das sequências de vídeo. O melhor resultado de cada vídeo é destacado em negrito.

Tabela 4.3 – Valor da métrica PosErr para os rastreadores avaliados no experimento.

Vídeo/Rastreador	TLD	FRAG	SCM	SURF
Face:bb	-	17,26	76,00	**12,74**
Face:po	19,56	5,09	**3,84**	10,16
Face:mb	-	9,89	**4,33**	10,39

Conforme pode ser observado na Tabela 4.3, o rastreador SCM foi superior aos demais nos vídeos Face:po e Face:mp e obteve o pior resultado no vídeo Face:bb, ao contrário do rastreador SURF proposto que obteve o melhor resultado neste último vídeo. De acordo com Wang et al. (2014), os demais valores do rastreador TLD não foram computados pois a região rastreada extrapolou as dimensões da imagem (*out-of-bounds*).

Na Tabela 4.4, são apresentados os resultados dos quatro rastreadores para a métrica F-score para cada uma das sequências de vídeo, destacados em negrito o melhor resultado de cada vídeo.

Tabela 4.4 – Valor da métrica F-score para os rastreadores avaliados no experimento.

Vídeo/Rastreador	TLD	FRAG	SCM	SURF
Face:bb	-	0,509	0,371	**0,621**
Face:po	0,783	0,946	**0,962**	0,873
Face:mb	-	0,727	**0,845**	0,714

Conforme pode ser observado na Tabela 4.4, novamente, o rastreador SCM foi superior aos demais nos vídeos Face:po e Face:mp e obteve o pior resultado no vídeo Face:bb, ao contrário do rastreador SURF proposto que obteve o melhor resultado neste último vídeo. De acordo com Wang et al. (2014), os demais valores do rastreador TLD não foram computados pois a região rastreada extrapolou as dimensões da imagem (*out-of-bounds*).

4.2.3. Conclusões

Apesar do rastreador SCM ter obtido melhor resultado no experimento realizado em duas das sequências de vídeos, não foi encontrada uma implementação sua na linguagem C/C++, adotada para a implementação do método proposto nesta Tese. O rastreador SURF proposto obteve o melhor resultado na sequência *Face:bb*. No entanto, os resultados nas outras duas sequências de vídeo foram inferiores aos dos rastreadores SCM e FRAG.

Desta forma, o rastreador FRAG foi selecionado para integração com o sistema proposto, uma vez que apresentou o segundo melhor resultado em todos os vídeos avaliados, além de possuir implementação na linguagem C++, facilitando o processo de integração com o sistema proposto. O rastreador TLD foi descartado, por não ter apresentado resultados suficientes para comparação.

4.3. Agrupamento de Faces em Vídeos

Nesta seção, são descritos e comentados os resultados obtidos a partir de três experimentos objetivos para avaliar desempenho do método de agrupamento de faces desenvolvido.

Tais experimentos objetivaram a avaliação da eficiência computacional em termos de velocidade de processamento (*performance*) e da qualidade do agrupamento resultante.

Na base de dados *YouTube Celebrities* (KIM et al., 2008), foram considerados os resultados publicados nos trabalhos de Kim et al. (2008), Hu et al. (2011), Cui et al. (2012) e Anoop et al. (2012). Na base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013), foram considerados os resultados publicados no trabalho de Anantharajah et al. (2015).

4.3.1. Base de Dados

A base de vídeos *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013), contém 55 vídeos de noticiários totalizando 123 minutos. Nesta base, estão presentes 110 pessoas e a mesma é subdividida em dois subconjuntos: *dev* e *eval*. O subconjunto *dev* consiste de 91.231 faces anotadas (em forma retangular) de 27 vídeos, em que 98 pessoas estão presentes (vide Figura 4.7 para alguns exemplos).

Figura 4.7 – Amostras de imagens da base *SAIVT-Bnews*.

Fonte: Adaptado de Ghaemmaghami, Dean e Sridharan (2013).

O subconjunto *eval* consiste de 81.812 faces anotadas de 28 vídeos, em que 26 pessoas estão presentes. O subconjunto *dev* é utilizado para ajustar os parâmetros dos sistemas, enquanto que o subconjunto *eval* é utilizado para avaliar o desempenho dos sistemas. Para informações acerca da base de dados *YouTube Celebrities* (KIM et al., 2008), vide Seção 4.1.1.

4.3.2. Resultados Preliminares

A partir da combinação de componentes internos, realizou-se um experimento preliminar a fim de mensurar a contribuição de cada componente no resultado final do sistema. Foram geradas 96 combinações de configuração utilizando os sete componentes que constituem o sistema, conforme detalhado a seguir:

- **F**iltragem **H**omomórfica: ligada ou desligada;
- **E**qualização de **H**istograma: ligada ou desligada;
- **D**etecção de ***S****hots*: ligada ou desligadoa;
- **D**etecção de **F**aces: *Cascade****CNN***, ***P****ICO* ou ***P****ICO* ***T****RAIN*;
- **R**astreamento de **F**aces: ***F****RAG* ou ***S****URF*;
- **A**grupamento de **F**aces: ligado ou desligado;
- **S**imilaridade **T**emporal / **R**eagrupamento **E**spacial: ligada ou desligada.

Para determinar qual a melhor combinação de componentes foram utilizadas as métricas *Average Purity* (pureza) e *Average Coverage* (cobertura). Adicionalmente, foi realizada a marcação manual de 30 vídeos escolhidos aleatoriamente da base de dados *YouTube Celebrities* (KIM et al., 2008) para a determinação do *ground-truth* da detecção de *shots* e de faces.

Na Tabela 4.5, são apresentados os resultados das duas métricas para cada possível combinação de componentes do método proposto, conforme a configuração do componente. A apresentação da tabela baseia-se na ordem descrescente das métricas *Average Purity* (pureza) e *Average Coverage* (cobertura).

Tabela 4.5 – Valores das métricas *Average Purity* e *Average Coverage* para cada combinação de componentes do método proposto avaliados no experimento intermediário.

#Comb	F.H	E.H	D.S	D.F	R.F	A.F	S.T/R.E	P_w	C_w
01	L	L	L	PT	F	L	L	0,8876	0,8611
02	D	L	L	PT	F	L	L	0,8734	0,8553
03	L	D	L	PT	F	L	L	0,8791	0,8574
04	L	L	D	PT	F	L	L	0,8138	0,8062
05	L	L	L	PT	F	L	D	0,8736	0,8581
06	D	D	L	PT	F	L	L	0,8682	0,8495
07	L	D	D	PT	F	L	L	0,8565	0,8397
08	L	L	D	PT	F	L	D	0,8482	0,8215
09	D	L	D	PT	F	L	L	0,8457	0,8201
10	D	L	L	PT	F	L	D	0,8333	0,8151
11	L	D	L	PT	F	L	D	0,8352	0,8159
12	D	D	D	PT	F	L	L	0,8209	0,8127
13	L	D	D	PT	F	L	D	0,8135	0,8062
14	D	L	D	PT	F	L	D	0,8168	0,8071
15	D	D	L	PT	F	L	D	0,8491	0,8366
16	D	D	D	PT	F	L	D	0,8046	0,7922
17	L	L	L	PT	S	L	L	0,8489	0,8275
18	D	L	L	PT	S	L	L	0,8472	0,8269
19	L	D	L	PT	S	L	L	0,8424	0,8215
20	L	L	D	PT	S	L	L	0,7832	0,7612
21	L	L	L	PT	S	L	D	0,8491	0,8277
22	D	D	L	PT	S	L	L	0,8308	0,8139
23	L	D	D	PT	S	L	L	0,8215	0,8076
24	L	L	D	PT	S	L	D	0,8168	0,7919
25	D	L	D	PT	S	L	L	0,8106	0,7894
26	D	L	L	PT	S	L	D	0,8037	0,7883
27	L	D	L	PT	S	L	D	0,8026	0,7867
28	D	D	D	PT	S	L	L	0,7991	0,7743
29	L	D	D	PT	S	L	D	0,7813	0,7612
30	D	L	D	PT	S	L	D	0,7816	0,7718
31	D	D	L	PT	S	L	D	0,8109	0,8023
32	D	D	D	PT	S	L	D	0,7704	0,7585
33	L	L	L	P	F	L	L	0,8469	0,8247
34	D	L	L	P	F	L	L	0,8453	0,8193

Tabela 4.5 – Valores das métricas *Average Purity* e *Average Coverage* para cada combinação de componentes do método proposto avaliados no experimento intermediário (continuação).

#Comb	F.H	E.H	D.S	D.F	R.F	A.F	S.T/R.E	P_w	C_w
35	L	D	L	P	F	L	L	0,8352	0,8145
36	L	L	D	P	F	L	L	0,8133	0,7946
37	L	L	L	P	F	L	D	0,8352	0,8189
38	D	D	L	P	F	L	L	0,8269	0,8014
39	L	D	D	P	F	L	L	0,8142	0,7971
40	L	L	D	P	F	L	D	0,8189	0,7988
41	D	L	D	P	F	L	L	0,8277	0,8025
42	D	L	L	P	F	L	D	0,8032	0,7832
43	L	D	L	P	F	L	D	0,8076	0,7834
44	D	D	D	P	F	L	L	0,7919	0,7703
45	L	D	D	P	F	L	D	0,7883	0,7691
46	D	L	D	P	F	L	D	0,7705	0,751
47	D	D	L	P	F	L	D	0,7722	0,7552
48	D	D	D	P	F	L	D	0,7481	0,7265
49	L	L	L	P	S	L	L	0,8327	0,8163
50	D	L	L	P	S	L	L	0,8139	0,7982
51	L	D	L	P	S	L	L	0,8167	0,7997
52	L	L	D	P	S	L	L	0,8023	0,7845
53	L	L	L	P	S	L	D	0,8272	0,8012
54	D	D	L	P	S	L	L	0,8104	0,7921
55	L	D	D	P	S	L	L	0,7919	0,7728
56	L	L	D	P	S	L	D	0,7867	0,7683
57	D	L	D	P	S	L	L	0,7894	0,7605
58	D	L	L	P	S	L	D	0,8096	0,7884
59	L	D	L	P	S	L	D	0,8039	0,7832
60	D	D	D	P	S	L	L	0,7667	0,7409
61	L	D	D	P	S	L	D	0,7468	0,7286
62	D	L	D	P	S	L	D	0,7492	0,7207
63	D	D	L	P	S	L	D	0,7718	0,7522
64	D	D	D	P	S	L	D	0,7305	0,7119
65	L	L	L	CNN	F	L	L	0,8523	0,8392
66	D	L	L	CNN	F	L	L	0,8493	0,8272
67	L	D	L	CNN	F	L	L	0,8475	0,8241
68	L	L	D	CNN	F	L	L	0,8201	0,8076

Tabela 4.5 – Valores das métricas *Average Purity* e *Average Coverage* para cada combinação de componentes do método proposto avaliados no experimento intermediário (continuação).

#Comb	F.H	E.H	D.S	D.F	R.F	A.F	S.T/R.E	P_w	C_w
69	L	L	L	CNN	F	L	D	0,8366	0,8139
70	D	D	L	CNN	F	L	L	0,8369	0,8167
71	L	D	D	CNN	F	L	L	0,8159	0,7894
72	L	L	D	CNN	F	L	D	0,8127	0,7881
73	D	L	D	CNN	F	L	L	0,8151	0,7883
74	D	L	L	CNN	F	L	D	0,8232	0,8023
75	L	D	L	CNN	F	L	D	0,8276	0,8039
76	D	D	D	CNN	F	L	L	0,8019	0,7805
77	L	D	D	CNN	F	L	D	0,7922	0,7718
78	D	L	D	CNN	F	L	D	0,7854	0,7697
79	D	D	L	CNN	F	L	D	0,7803	0,7624
80	D	D	D	CNN	F	L	D	0,7618	0,7465
81	L	L	L	CNN	S	L	L	0,8402	0,8269
82	D	L	L	CNN	S	L	L	0,8395	0,8168
83	L	D	L	CNN	S	L	L	0,8352	0,8139
84	L	L	D	CNN	S	L	L	0,8064	0,7883
85	L	L	L	CNN	S	L	D	0,8275	0,8023
86	D	D	L	CNN	S	L	L	0,8139	0,7919
87	L	D	D	CNN	S	L	L	0,8071	0,7883
88	L	L	D	CNN	S	L	D	0,8062	0,7867
89	D	L	D	CNN	S	L	L	0,8153	0,7922
90	D	L	L	CNN	S	L	D	0,7989	0,7722
91	L	D	L	CNN	S	L	D	0,7987	0,7705
92	D	D	D	CNN	S	L	L	0,7881	0,7647
93	L	D	D	CNN	S	L	D	0,7743	0,7519
94	D	L	D	CNN	S	L	D	0,7612	0,7408
95	D	D	L	CNN	S	L	D	0,7585	0,7311
96	D	D	D	CNN	S	L	D	0,7384	0,7153

Conforme pode ser observado na Tabela 4.5, à medida que os componentes são desligados, os valores das métricas P_w e C_w decrescem consideravelmente, independentemente do tipo de detector e rastreador de faces. Outro fator importante é a contribuição positiva ao considerar os componentes de detecção de *shots* e similaridade temporal / reagrupamento

espacial no *pipeline* do sistema. Similarmente, pode-se observar que a melhor combinação de métodos para o sistema considera o detector de faces PICO TRAIN e o rastreador de faces FRAG corroborando com os resultados obtidos nos experimentos detalhados nas seções anteriores.

Adicionalmente, a partir da melhor combinação de componentes determinada experimentalmente, foram computados os valores de acurácia para o detector de *shots*, o detector de faces e o rastreador de faces aliadas aos respectivos valores das métricas P_w e C_w no sistema final, afim de mensurar suas acurácias individualmente (vide Tabela 4.6).

Tabela 4.6 – Valores de TPR, P_w e C_w da melhor combinação de componentes do método proposto.

-	D.S	D.F	R.F	P_w	C_w
ACC	0,9384	0,9503	0,9257	0,8876	0,8611

4.3.3. Avaliação Comparativa

Os métodos comparados no experimento da base de dados *YouTube Celebrities* (KIM et al., 2008), foram *Visual Constraints* (KIM et al., 2008), *Sparse Approximated Nearest Points* (HU et al., 2011), *Image Sets Alignment* (CUI et al., 2012), *Principal Angles, Kernelized Principal Angles e Covariance Profiles* (ANOOP et al., 2012), e o método proposto. Para a base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013) foram comparados os métodos *Local GMM* e *Local TVM* (ANANTHARAJAH et al., 2015), e o método proposto.

Para determinar qual o melhor método de agrupamento de faces em vídeos foi utilizado as métricas *Average Purity* (pureza) e *Average Coverage* (cobertura). Para obter estas métricas, cada grupo é analisado e rotulado com a identidade de face (pessoa) de maior frequência. A pureza (*P*) de um grupo, pode ser calculada como (GHAEMMAGHAMI et al., 2012):

$$P = \frac{N^i}{N_t^i}, \tag{4.7}$$

em que N^i representa o número de faces rotuladas no *i*-ésimo grupo e N_t^i representa o número total de faces presentes no *i*-ésimo grupo. Para cada pessoa *j*, o grupo contendo o maior número de faces da pessoa *j*, $max(N_j)$ é

calculado. A métrica *Coverage* (*C*), pode ser calculada como: (GHAEMMAGHAMI et al., 2012):

$$C = \frac{\max(N_j)}{N_t^j}, \tag{4.8}$$

em que N_t^j denota o número total de faces da pessoa *j* disponível no vídeo de acordo com a anotação manual. Assim, os valores da *Average Purity* (P_w) e da *Average Coverage* (C_w) utilizados para avaliar o desempenho de métodos de agrupamento são calculadas como (GHAEMMAGHAMI et al., 2012):

$$P_w = \frac{\sum_{t=1}^{t=F} N_t \times P_t}{\sum_{t=1}^{t=F} N_t}, \tag{4.9}$$

em que P_t denota o *t*-ésimo grupo formado pelas faces detectadas do vídeo, N_t representa o total de faces presentes no grupo *t* e *F* é o número total de grupos formados pelo método de agrupamento. O valor de *Average Coverage* (C_w) é calculado conforme a Equação (4.10) (GHAEMMAGHAMI et al., 2012):

$$C_w = \frac{\sum_{s=1}^{s=M} R_s \times C_s}{\sum_{s=1}^{s=M} R_s}, \tag{4.10}$$

em que C_s denota a cobertura da *s*-ésima pessoa presente no vídeo, R_s representa a quantidade total de faces da *s*-ésima pessoa de acordo com o *ground-truth* e *M* é o número de pessoas presentes no vídeo. Ambas as métricas variam no intervalo [0,1].

Apenas para o método proposto, além das duas métricas P_w e C_w, os valores de sete diferentes métricas foram computados no experimento (vide Apêndice A): (i) *Rand Index* (RI); (ii) *Adjusted Rand Index* (ARI); (iii) *Precision* (P); (iv) *Recall* (R); (v) *F-Mesure* (F); (vi) *Jaccard Index* (JI); (vii) *Folkes and Mallows Index* (FMI). Tais métricas assumem valores no intervalo [0,1], em que valores próximos de 1 indicam melhor resultado.

Na Tabela 4.7, são apresentados os resultados do método de

agrupamento proposto para cada métrica na base de dados *YouTube Celebrities* (KIM et al., 2008).

Tabela 4.7 – Resultados (média, variância e desvio-padrão) das métricas de avaliação de agrupamento para o método proposto na base na base de dados *YouTube Celebrities* (KIM et al., 2008).

-	RI	ARI	P	R	F	JI	FMI	P_w	C_w
μ	0,9276	0,8557	0,9417	0,8240	0,8688	0,8160	0,8748	0,8920	0,8847
σ^2	0,0041	0,0131	0,0025	0,0219	0,0107	0,0224	0,0093	0,0063	0,0074
σ	0,0638	0,1147	0,0501	0,1480	0,1035	0,1498	0,0965	0,0795	0,0863

Em seguida, foi realizado um teste visual de Intervalo de Confiança (IC) conhecido por *boxplot* (JAIN, 1991) sobre os dados de cada métrica de avaliação de agrupamento. Testes visuais são utilizados para verificar graficamente o comportamento dos dados obtidos, i.e., avaliar a simetria dos dados, sua dispersão e a existência ou não de *outliers*, sendo especialmente adequados para a comparação de dois ou mais conjuntos de dados correspondentes às categorias de uma variável qualitativa (LEVINE, BERENSON e STEPHAN, 2000).

Na Figura 4.8, são exibidos os gráficos *boxplot*, para os dados obtidos de cada uma das sete métricas consideradas. Uma descrição sucinta da construção e da utilização dos gráficos *boxplot* encontra-se no Apêndice B.

Figura 4.8 – Gráficos boxplot das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) Pw; e (J) Cw.

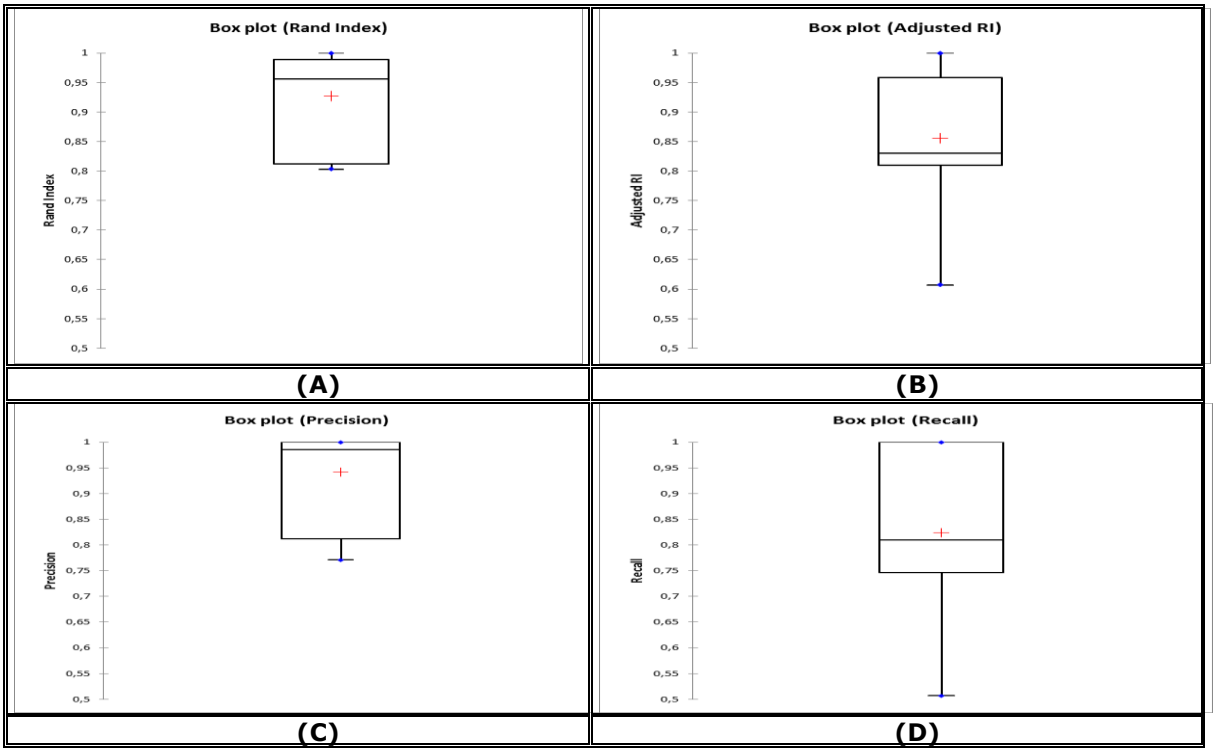
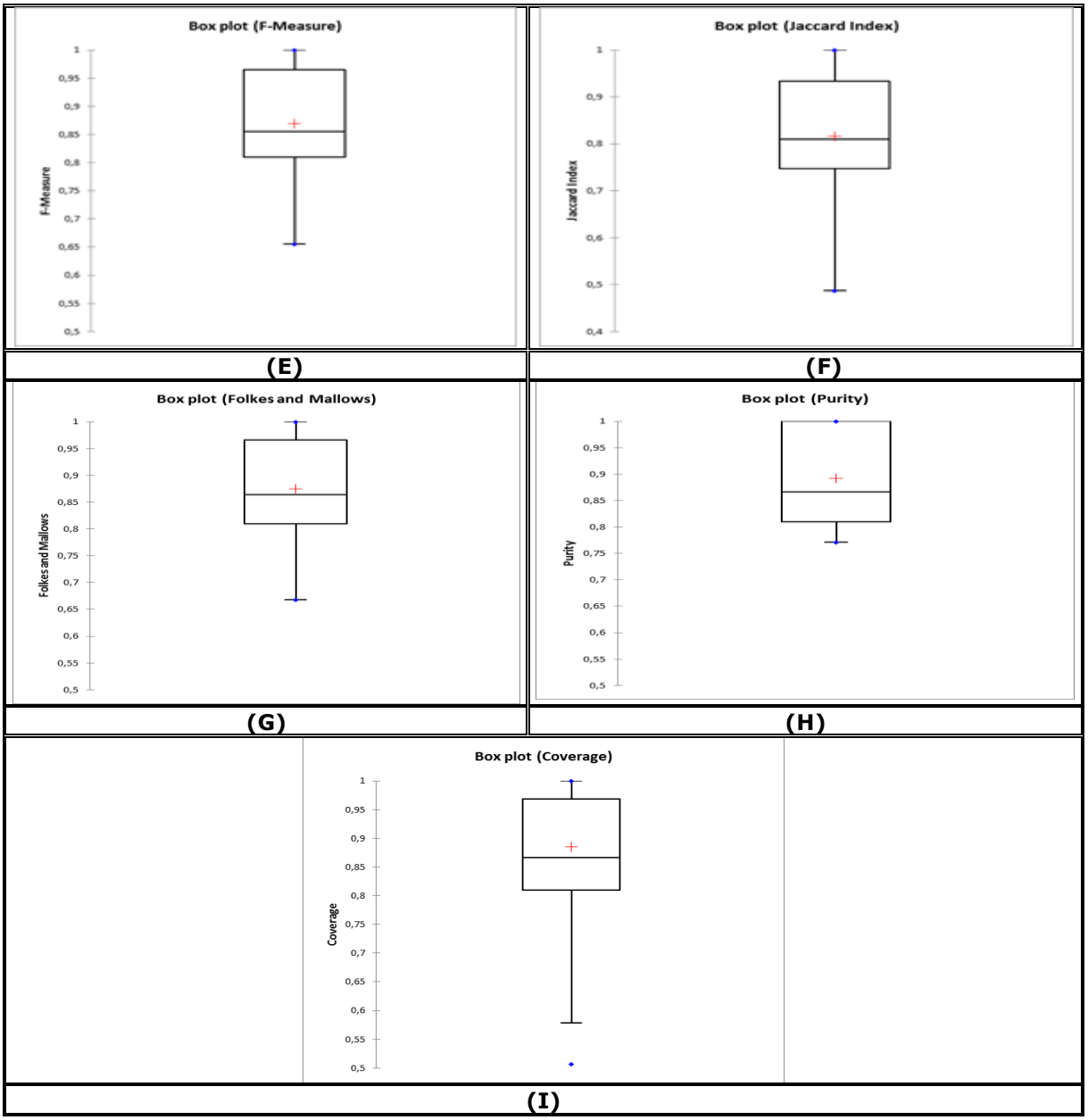


Figura 4.8 – Gráficos *boxplot* das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) P_w; e (J) C_w (continuação).



Analisando os gráficos *boxplot* da Figura 4.8, foi verificado que as métricas RI, P e P_w apresentam a menor variação nos dados, demonstrando elevado grau de qualidade do método proposto. Adicionalmente, a partir dos gráficos *boxplot* das métricas R, JI e C_w, pode-se perceber que estas foram afetadas pelo resultado negativo de *outliers*, prejudicando de certa maneira o desempenho do método de agrupamento proposto na base de dados *YouTube Celebrities* (KIM et al., 2008).

Na Tabela 4.8, são apresentados os resultados comparativos dos métodos de agrupamento avaliados para a métrica *Average Purity* (P_w) na base de dados *YouTube Celebrities* (KIM et al., 2008). Os autores dos

trabalhos comparados não forneceram resultados para a métrica *Average Coverage* (C_w) nessa base.

Conforme pode ser observado na Tabela 4.8, o método proposto obteve o melhor resultado em comparação com os demais trabalhos analisados. Adicionalmente, com base nos resultados apresentados na Tabela 4.6, pode-se observar que, para todas as métricas, o valor obtido pelo método proposto foi superior a 0,8 indicando que pode ser considerado um resultado satisfatório dado que está próximo do valor 1,0 (cenário ideal).

Tabela 4.8 – Valor da métrica *Purity* para os métodos de agrupamento avaliados na base *YouTube Celebrities* (KIM et al., 2008).

Método	P_w
Sparse Approximated Nearest Points – Hu et al. (2011)	0,6503
Visual Constraints – Kim et al. (2008)	0,7000
Image Sets Alignment – Cui et al. (2012)	0,7460
Principal Angles – Anoop et al. (2012)	0,7860
Kernelized Principal Angles – Anoop et al. (2012)	0,7970
Covariance Profiles - Anoop et al. (2012)	0,8035
Proposto	**0,8920**

No segundo experimento em que foi considerada a base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013), foram comparados os métodos *Local GMM* e *Local TVM* (ANANTHARAJAH et al., 2015) e o método proposto.

Novamente, apenas para o método proposto, foram calculadas todas as nove métricas de avaliação de agrupamento. Na Tabela 4.9, são apresentados os resultados do método proposto para cada métrica no subconjunto *dev*.

Tabela 4.9 – Resultados (média, variância e desvio-padrão) das métricas de avaliação de agrupamento para o método proposto no subconjunto *dev* da base de dados *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

-	RI	ARI	P	R	F	JI	FMI	P_w	C_w
μ	0,9609	0,8896	0,9848	0,8933	0,9209	0,8533	0,8853	0,9563	0,9448
σ^2	0,0013	0,0015	0,0026	0,0031	0,0021	0,0054	0,0034	0,0008	0,0005
σ	0,0362	0,0390	0,0510	0,0556	0,0455	0,0735	0,0583	0,0274	0,0215

Em seguida, foi realizado um teste visual de Intervalo de Confiança (IC) conhecido por *boxplot* (JAIN, 1991) sobre os dados de cada métrica de avaliação de agrupamento do subconjunto *dev*, conforme ilustrado na Figura 4.9.

Figura 4.9 – Gráficos *boxplot* das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) P_w; e (J) C_w.

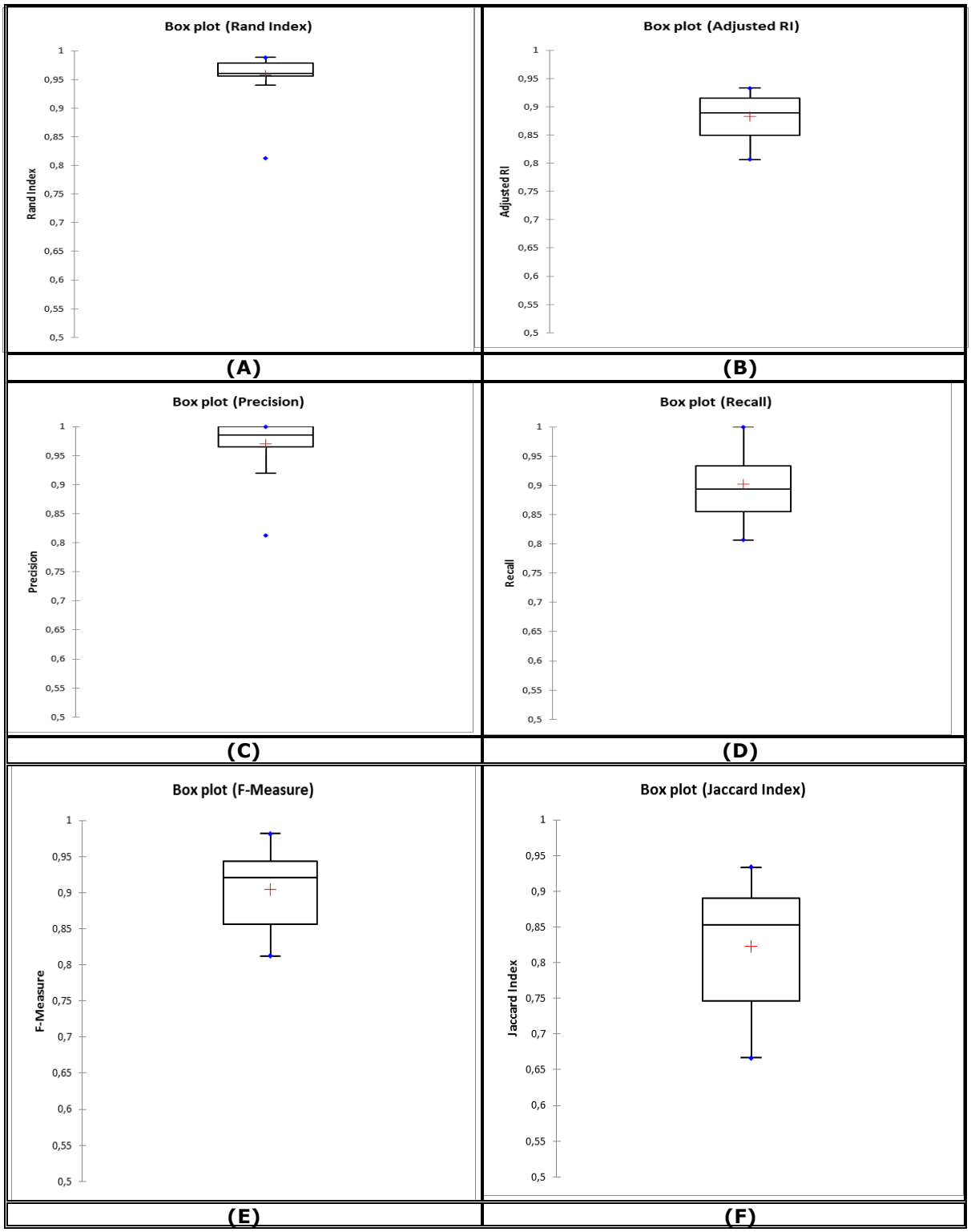
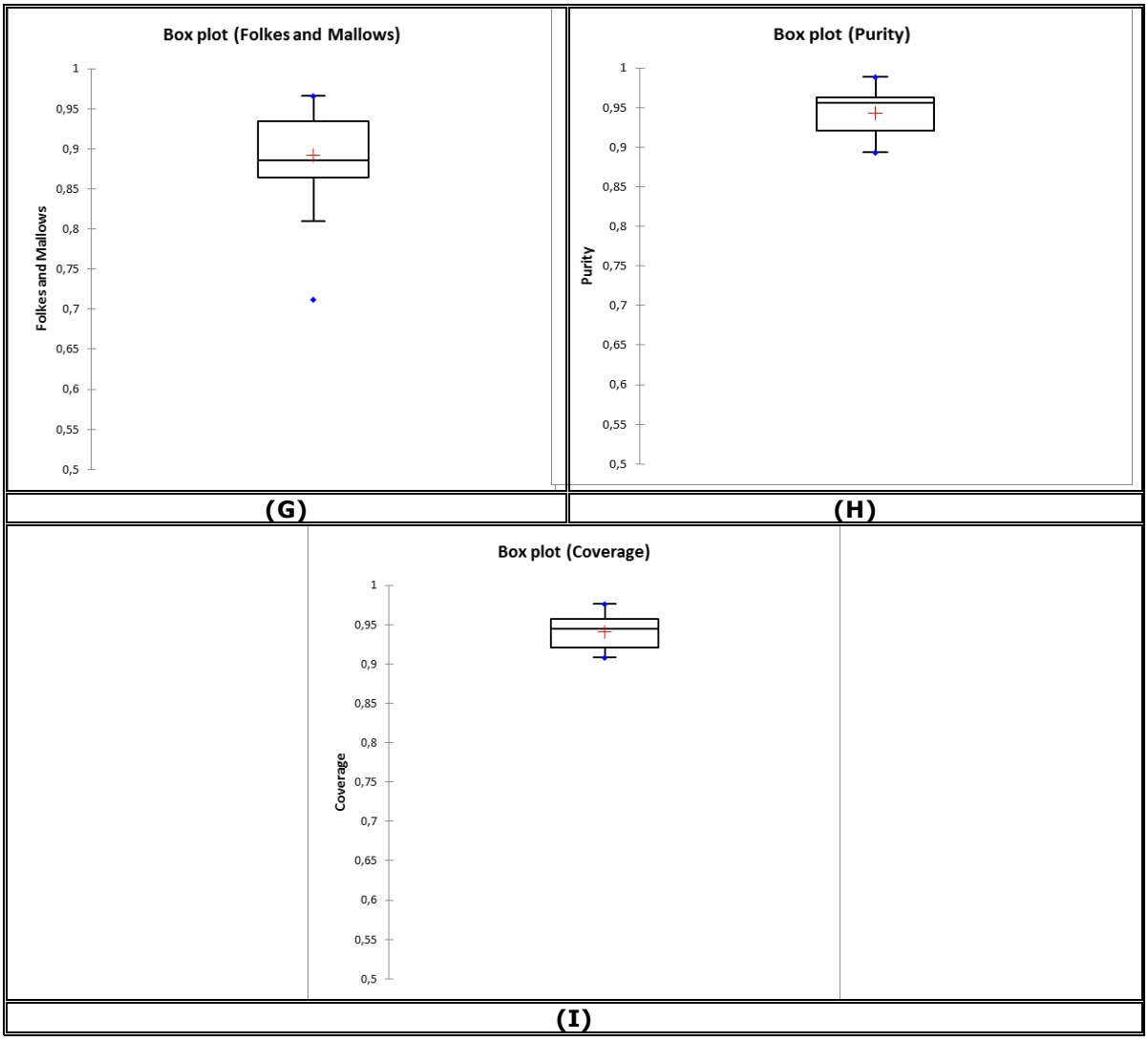


Figura 4.9 – Gráficos *boxplot* das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) P_w; e (J) C_w (continuação).



Analisando os gráficos *boxplot* da Figura 4.9, foi verificado que as métricas RI, P, P_w e C_w apresentam a menor variação nos dados, demonstrando um elevado grau de qualidade do método proposto. Na Tabela 4.10 são apresentados os resultados comparativos dos métodos de agrupamento avaliados para as métricas *Average Purity* (P_w) e *Average Coverage* (C_w) no subconjunto *dev* da base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Tabela 4.10 – Valor da métrica *Purity* e *Coverage* para os métodos de agrupamento avaliados no subconjunto *dev* da base de dados *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Método	P_w	C_w
Local GMM – Anantharajah et al. (2015)	0,9840	0,2250
Local TVM – Anantharajah et al. (2015)	0,9888	0,7090
Proposto	0,9875	**0,8367**

Considerando-se que os valores da métrica P_w são muito próximos, pode-se perceber que o método proposto obteve o melhor resultado em comparação com os demais trabalhos analisados (melhor valor de C_w), conforme a Tabela 4.10.

Adicionalmente, com base nos resultados apresentados na Tabela 4.9, pode-se observar que, para todas as métricas, o valor obtido pelo método proposto foi superior a 0,85 indicando que pode ser considerado um resultado satisfatório dado que está próximo do valor 1,0 (cenário ideal).

O mesmo procedimento descrito anteriormente foi realizado para o subconjunto *eval* da base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013). Na Tabela 4.11 são apresentados os resultados do método proposto para cada métrica no subconjunto *eval*.

Tabela 4.11 – Resultados das métricas de avaliação de agrupamento para o método proposto no subconjunto *eval* da base de dados *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

-	RI	ARI	P	R	F	JI	FMI	P_w	C_w
μ	0,9609	0,9059	0,9848	0,8933	0,9209	0,8733	0,8933	0,9609	0,9520
σ^2	0,0007	0,0098	0,0016	0,0040	0,0017	0,0039	0,0059	0,0005	0,0005
σ	0,0270	0,0992	0,0396	0,0629	0,0416	0,0620	0,0769	0,0220	0,0232

Em seguida, foi realizado um teste visual de Intervalo de Confiança (IC) conhecido por *boxplot* (JAIN, 1991) sobre os dados de cada métrica de avaliação de agrupamento do subconjunto *eval*, conforme ilustrado na Figura 4.10.

Figura 4.10 – Gráficos *boxplot* das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) P_w; e (J) C_w.

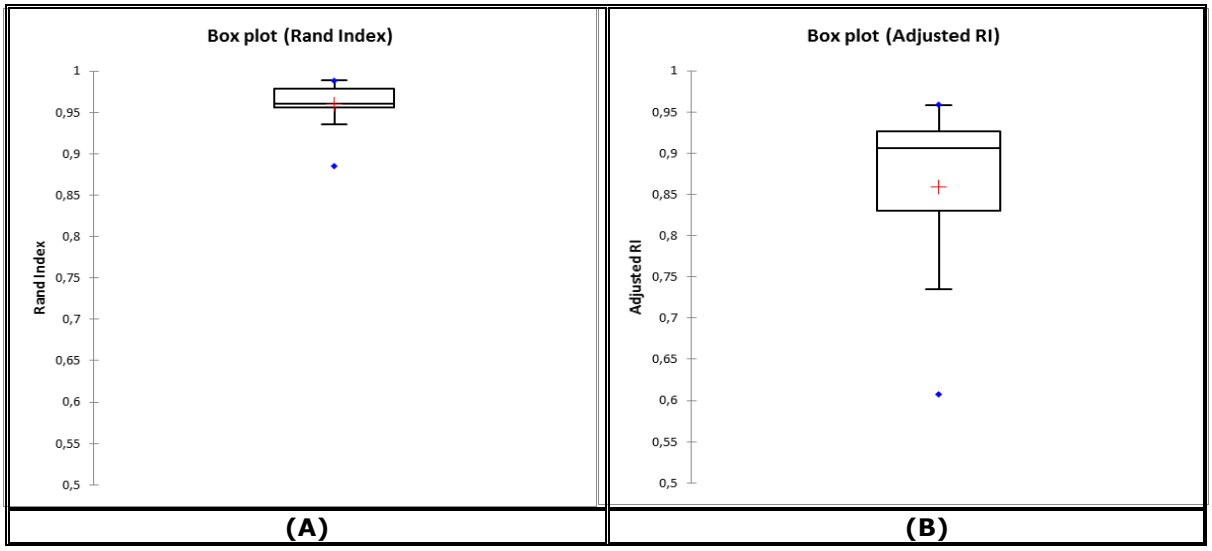
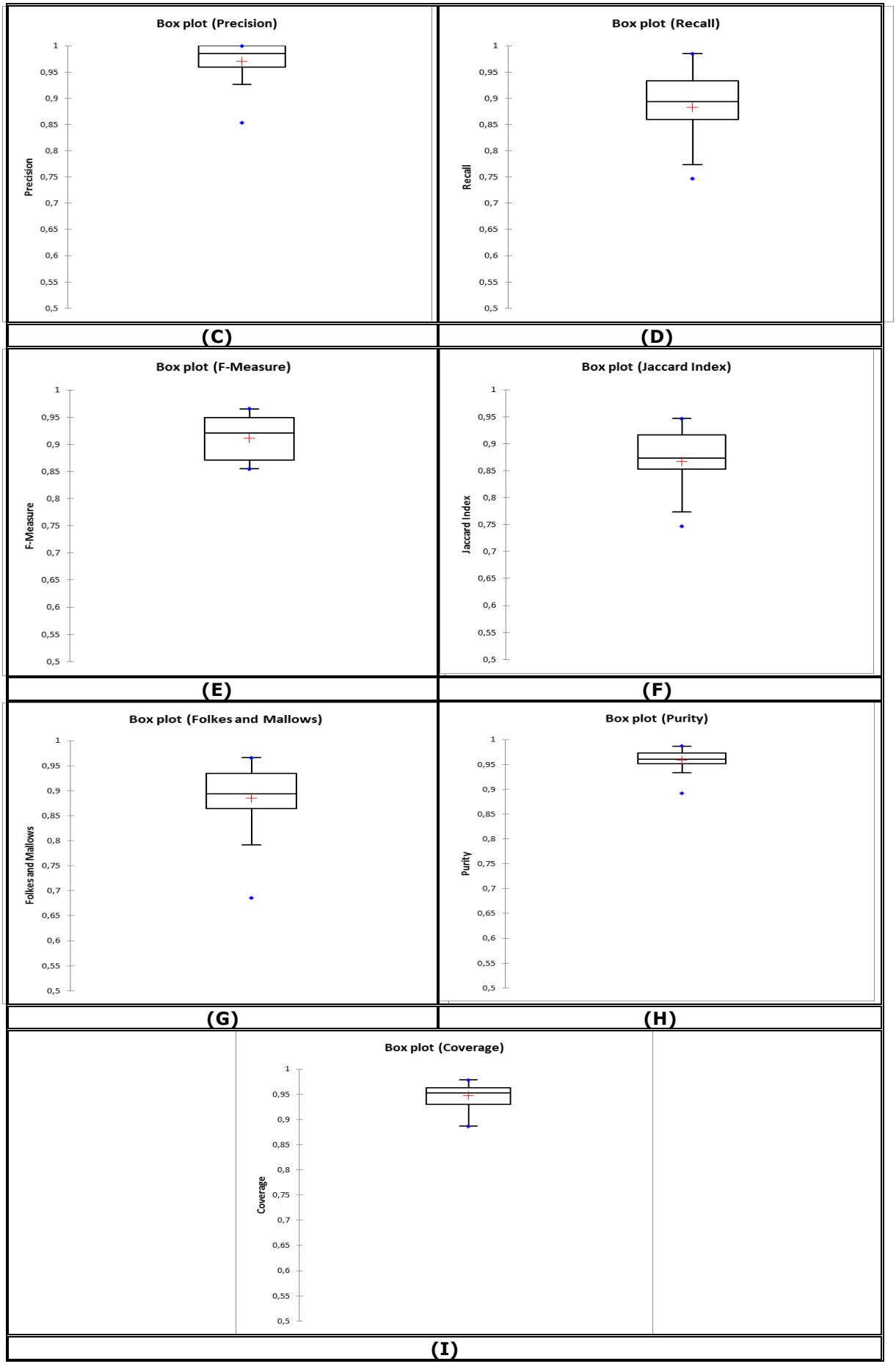


Figura 4.10 – Gráficos *boxplot* das métricas de avaliação de agrupamento: (A) RI; (B) ARI; (C) P; (D) R; (E) F; (F) JI; (G) FM; (H) P_w; e (J) C_w (continuação).



Analisando-se os gráficos *boxplot* da Figura 4.10, foi verificado que, assim como no conjunto *dev*, as métricas RI, P, P_w e C_w apresentam a menor variação nos dados, demonstrando um elevado grau de qualidade do método proposto.

Na Tabela 4.12, são apresentados os resultados comparativos dos métodos de agrupamento avaliados para as métricas *Average Purity* (P_w) e *Average Coverage* (C_w) no subconjunto *eval* da base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Tabela 4.12 – Valor da métrica *Purity* e *Coverage* para os métodos de agrupamento avaliados no subconjunto *eval* da base de dados *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Método	P_w	C_w
Local GMM – Anantharajah et al. (2015)	0,9920	0,1920
Local TVM – Anantharajah et al. (2015)	0,9810	0,7700
Proposto	0,9907	**0,8492**

Considerando-se que os valores da métrica P_w são equivalentes, pode-se perceber que o método proposto obteve o melhor resultado em comparação com os demais trabalhos analisados (melhor valor de C_w), conforme a Tabela 4.12.

Adicionalmente, com base nos resultados apresentados na Tabela 4.11, pode-se observar que para todas as métricas o valor obtido pelo método proposto foi superior a 0,87, indicando que pode ser considerado um resultado satisfatório dado que está próximo do valor 1,0 (cenário ideal).

Outro fator importante para a comparação dos métodos é o tempo médio de processamento, seja no tempo total ou o tempo gasto em cada etapa de processamento. Os tempos de processamento do método proposto foram obtidos considerando uma CPU de 2.6 GHz *dual core* com 8 GB de memória RAM, com o sistema operacional *Windows 7*.

Apenas o estudo de Hu et al. (2011) apresentou o tempo médio de execução na base *YouTube Celebrities* (KIM et al., 2008), conforme representado na Tabela 4.13. O tempo médio de um vídeo nessa base é de 6,0523 segundos de duração.

Tabela 4.13 – Tempo médio de execução de cada método na base *YouTube Celebrities* (KIM et al., 2008).

Método	Tempo (s)
Sparse Approximated Nearest Points – Hu et al. (2011)	55,64
Proposto	**19,17**

Pelo resultado apresentado na Tabela 4.13, evidencia-se que o método proposto é superior ao método de Hu et al. (2011), no tocante ao tempo total de processamento. Na Tabela 4.14, são apresentados os tempos de processamento de cada componente do método proposto (tempo total de 19,17 segundos) na base *YouTube Celebrities* (KIM et al., 2008):

Tabela 4.14 – Valor da métrica *Purity* e *Coverage* para os métodos de agrupamento avaliados no subconjunto *eval* da base de dados *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Método	Tempo (s)
Extração e Correção de Quadros	1,78
Detecção de *Shots*	1,13
Detecção e Rastreamento de Faces	5,84
Extração de Características e Agrupamento de Faces	5,69
Verificação de Similaridade Temporal	2,35
Reagrupamento Espacial	2,38

No estudo de Anantharajah et al. (2015) foram calculados tempos para comparação e *enrollment* entre os *face tracklets* com base na quantidade de faces representativas a serem comparadas, considerando a base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Assim, no método proposto, foi contabilizada apenas a parcela referente ao tempo de comparação de *face tracklets*, conforme representado na Tabela 4.15. O tempo médio de um vídeo nessa base é de 130,4736 segundos de duração.

Tabela 4.15 – Tempo médio (em segundos) de execução de cada método na base *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

Faces Representativas	Local TVM Anantharajah et al. (2015)	Proposto
1	0,03	0,01
2	0,03	0,01
5	0,04	0,02
10	0,04	0,04
20	0,06	0,08

Conforme os dados da Tabela 4.15, pode-se considerar também que o método proposto na maioria dos cenários de número de faces representativas é mais rápido que o método de Anantharajah et al. (2015) na etapa de comparação de face *tracklets*, exceto no cenário de 20 faces representativas.

4.3.4. Conclusões

Em função dos valores numéricos apresentados pela avaliação do método proposto nas bases *YouTube Celebrities* (KIM et al., 2008) e *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013), evidenciou-se sua superioridade em relação aos demais métodos comparados. No tocante aos aspectos de qualidade de agrupamento segundo métricas objetivas de avaliação de agrupamento.

Adicionalmente, em relação ao tempo de processamento, novamente, o método proposto foi superior aos métodos concorrentes nas bases *YouTube Celebrities* (KIM et al., 2008) e *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

4.4. Considerações Finais

Neste capítulo, foi descrita a parte experimental do estudo desenvolvido nesta tese. Tal parte teve como propósito a validação de componentes que constituem a abordagem proposta e descrita no Capítulo 3. Tais componentes e métodos foram avaliados por meio de testes visuais (análise gráfica) e testes numéricos, em comparação com outros resultados provenientes de trabalhos relacionados.

A partir da análise dos resultados obtidos nos experimentos objetivos realizados, pôde-se concluir que existe suporte numérico (métricas de avaliação de qualidade de agrupamento) para afirmar que a abordagem proposta apresenta desempenho superior para a tarefa de agrupamento de faces em vídeos digitais.

A avaliação experimental realizada, composta de testes visuais, testes de desempenho com base em métricas de avaliação de agrupamento,

poderá ser utilizada como base para experimentos futuros com outras ferramentas e abordagens para o problema objeto de estudo.

No próximo capítulo, é apresentada uma síntese da pesquisa ora reportada. Além disto, são apresentadas as contribuições e formuladas propostas para trabalhos futuros, as quais levam em conta as dificuldades e os problemas encontrados durante o desenrolar da pesquisa.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste capítulo, são apresentadas: (i) uma síntese dos principais tópicos abordados nesta Tese de Doutorado; (ii) as contribuições obtidas e esperadas; e (iii) proposições para pesquisas futuras, levando-se em conta as dificuldades e os problemas encontrados durante o desenrolar da pesquisa.

5.1. Síntese da Pesquisa

A pesquisa desenvolvida nesta Tese de Doutorado teve como meta central a concepção de uma abordagem inovadora para a solução do problema de agrupamento de faces em vídeos digitais. Como motivação, tem-se o crescente interesse nesta área pela indústria de tecnologias da informação, assim como pela lacuna de pesquisas sobre este tema, fundamentados em abordagens bem documentadas, com avaliações experimentais e estatísticas e que apresentem um arcabouço bem definido, visando *benchmarking*. O preenchimento desta lacuna, sob a forma de uma revisão sistemática da área, é uma das contribuições desta proposta, conforme destacado na próxima seção.

Durante o desenvolvimento da Tese de Doutorado, foram conduzidos estudos relacionados a cada um dos principais tópicos identificados para o problema investigado: (i) taxonomia da área; (ii) características extraídas para a identificação pessoal; (iii) métricas de similaridade; (iv) técnicas de agrupamento/reconhecimento; e (v) métricas de avaliação de agrupamento. Tais estudos foram apresentados no Capítulo 2 desta dissertação e serviram como fundamentação para a concepção da arquitetura da abordagem proposta. Além disto, os referidos estudos também fundamentaram a

implementação de componentes de *software* utilizados no processo de validação desta abordagem.

Com base nos estudos iniciais, foi desenvolvida uma abordagem original, fundamentada em técnicas e métodos do estado-da-arte em agrupamento de faces, relatado no Capítulo 2. Os detalhes desta extensão, da arquitetura, dos módulos, das técnicas e dos algoritmos que compõem a abordagem proposta foram apresentados no Capítulo 3 e também no Capítulo 4.

Os componentes integrantes da abordagem proposta foram submetidos a experimentos para a avaliação de sua viabilidade. Os resultados obtidos foram comparados com aqueles documentados em estudos relevantes da área de agrupamento de faces em vídeos digitais. Estes resultados, juntamente com a descrição de todos os procedimentos experimentais utilizados, foram apresentados no Capítulo 4.

Por fim, foi possível concluir, a partir dos experimentos objetivos conduzidos e do processamento estatístico dos dados coletados via *boxplots*, que os resultados produzidos pela abordagem proposta apresentaram melhores taxas na maioria das métricas de avaliação de agrupamento adotadas, quando comparadas aos resultados equivalentes produzidos por pesquisas relacionadas. Considerações acerca desta avaliação e conclusões são apresentadas na próxima seção.

5.2. Contribuições

A partir dos resultados experimentais apresentados no Capítulo 4, pode-se concluir que o objetivo geral desta Tese de Doutorado foi atingido, ou seja, o desenvolvimento de uma abordagem inovadora e robusta para o problema de agrupamento de faces em vídeos, visando à obtenção de um melhor desempenho em relação ao estado-da-arte. Para validar os componentes integrantes da abordagem proposta, um estudo experimental com testes visuais e numéricos foi conduzido em base de vídeos de referência, tais como, *YouTube Celebrities* (KIM et al., 2008) e *SAIVT-Bnews* (GHAEMMAGHAMI, DEAN e SRIDHARAN, 2013).

A abordagem proposta tem como principal inovação a agregação de módulos para atenuar os efeitos da queda de desempenho de agrupamento, normalmente associada a variações de iluminação, expressões faciais e pose. As principais contribuições da pesquisa podem ser enumeradas como segue:

(1) Execução de uma revisão bibliográfica sobre o estado-da-arte em Agrupamento de Faces em Vídeos, e avaliação de qualidade dos estudos relacionados, no tocante ao problema objeto de estudo, com base em dados coletados de bibliotecas digitais de referência da área;

(2) Concepção de uma taxonomia do corpo bibliográfico relacionado às principais abordagens utilizadas para o problema de agrupamento de faces. Os estudos selecionados foram categorizados em dois grupos dependendo de quais propriedades do vídeo exploram. *Abordagens com base em conjuntos de quadros* tratam os vídeos como coleções desordenadas de imagens e aproveitam a multiplicidade de observações, enquanto que as *abordagens com base em sequências de quadros*, explicitamente, utilizam a informação temporal para aumentar a eficiência ou permitir o reconhecimento em condições adversas;

(3) Proposição de método fundamentado em características SURF para as tarefas de rastreamento e agrupamento de faces, dado que nenhum trabalho relacionado utilizou esta característica para o problema em questão;

(4) Concepção de abordagem que permite a colaboração entre rastreamento e agrupamento de faces com base nos grupos formados após a vinculação dos face *tracklets* dando origem as faces representativas. Tal atividade torna-se necessária devido à grande variabilidade de faces (ocasionadas por mudanças de iluminação, pose, oclusão, etc.) que tendem a heterogeneizar o grupo, diminuindo sua similaridade intergrupo afetando negativamente o processo de agrupamento final por meio de um

agrupamento aglomerativo hierárquico espaço-temporal.

A seguir, são apresentadas sugestões para a melhoria e/ou para a complementação dos resultados obtidos nesta pesquisa.

5.3. Pesquisas Futuras

Nesta seção, apresentam-se algumas sugestões para pesquisas futuras, visando-se à obtenção de novas funcionalidades do sistema proposto. Algumas técnicas e métodos adotados na abordagem proposta podem ser substituídas por algoritmos similares em cada um dos módulos propostos, e.g., *Gradientfaces* (ZHANG et al., 2009) para compensação da iluminação; uso de características SIFT ou LDA, ao invés do SURF; algoritmos de agrupamento *K-Means*, EM, dentre outros, ao invés do HAC.

Portanto, propõe-se como pesquisa futura investigar a substituição de alguns componentes principais da abordagem para a verificação de possível melhora no desempenho do sistema.

Outra proposta para pesquisa futura é a adoção de outras bases de vídeos, por exemplo, *YouTube Faces* (WOLF et al., 2011), por apresentarem grande quantidade variações de indivíduos, expressões faciais e pose, com o objetivo de avaliar o comportamento do sistema e das demais abordagens concorrentes neste cenário, tais como, Otto, Wang e Jain (2016), Schroff, Kalenichenko e Philbin (2015), Zhou et al. (2015) e Bhatt et al. (2014).

Referências Bibliográficas

ADAM, A.; RIVLIN, E.; SHIMSHONI, I. Robust Fragments-based Tracking using the Integral Histogram. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2006.

ANANTHARAJAH, K.; DENMAN, S.; TJONDRONEGORO, D.; SRIDHARAN, S.; FOOKES, C. Robust Automatic Face Clustering in News Video. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Adelaide, SA, pp. 1-8, 2015.

ANOOP, K. R.; MITRA, A.; BONDE, U.; BHATTACHARYYA C.; RAMAKRISHNAN, K. R. Covariance profiles: A signature representation for object sets. *International Conference on Pattern Recognition (ICPR)*, Tsukuba, pp. 2541-2544, 2012.

ANTONOPOULOS, P.; NIKOLAIDIS, N; PITAS, I. Hierarchical Face Clustering using SIFT Image Features. *IEEE Symposium on Computational Intelligence in Image and Signal Processing*, 325-329, 2007.

APOSTOLIDIS, E.; MEZARIS, V. Fast shot segmentation combining global and local visual descriptors. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, pp. 6583-6587, 2014.

APOSTOLOFF, N. E.; ZISSERMAN, A. Who Are You? -- Real-time Person Identification. *British Machine Vision Conference*, 2007.

BAILLY-BAILLIÉRE, E. ET AL. The BANCA database and evaluation protocol. *In Proceeding of the Audio and Video-Based Biometric Person Authentication*, Springer-Verlag, 625–638, 2003.

BAUER, J.; SÜNDERHAUF, N; PROTZEL, P. Comparing Several Implementations of Two Recently Published Feature Detectors. *In Proceedings of the International Conference on Intelligent and Autonomous Systems* (IAV), Vol. 22, 481-494, 2007.

BAUML, M.; ROTH, M.; NEVATIA, R.; STIEFELHAGEN, R. Robust multi-pose face tracking by multi-stage tracklet association. In *International Conference on Pattern Recognition (ICPR)*, 2013.

BAY, H.; TUYTELAARS, T.; VAN GOOL, L. Surf: Speeded up robust features. *In Proceedings of the European Conference on Computer Vision* (ECCV), 404-417, 2006.

BELLHUMER, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. Fisherfaces: Recognition Using Specific Linear Projection. *European Conference on Computer Vision* (ECCV), 45-58, 1996.

BHATT, H. S.; SINGH, R.; VATSA, M. On rank aggregation for face recognition from videos. *In Proceedings of the 20th IEEE ICIP*, 1–5, 2014.

BISHOP, C. Pattern recognition and machine learning. New York: *Springer*, 2006.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation

of machine learning algorithms. *Pattern Recognition,* 30 (7), 1145-1159, 1997.

CALTECH FACE DATABASE. Disponível em: <http://www.vision.caltech.edu/archive.html>, 1999. Acesso em 27/12/2010.

CAO, X.; ZHANG, C.; FU, H.; LIU, S.; ZHANG, H. Diversity-induced Multi-view Subspace Clustering. *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Boston, MA, pp. 586-594, 2015A.

CAO, X.; ZHANG, C.; ZHOU, C.; FU H.; FOROOSH, H. Constrained Multi-View Video Face Clustering. *In IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4381-4393, 2015B.

CEVIKALP, H.; TRIGGS, B. Face recognition based on image sets. *In Proceeding of the International Conference on Computer Vision and Pattern Recognition* (CVPR), 2567–2573, 2010.

CHEN, Y. C.; PATEL, V. M.; PHILLIPS, P. J.; CHELLAPPA, R. Dictionary-based face recognition from video. *In Proceedings of the European Conference on Computer Vision* (ECCV), 766–779, 2012.

CHOI, J. Y.; DE NEVE, W.; RO, Y. M. Towards an Automatic Face Indexing System for Actor-based Video Services in an IPTV Environment. *In Proceedings of the IEEE Transactions on Consumer Electronics*, 147-155, 2010.

COOTES, T. F.; TAYLOR, C. J.; COOPER, D. H.; GRAHAM, J. Active Shape Models - Their Training and Application, *Computer Vision and Image Understanding*, Vol. 61, 38-59, 1995.

CUI, J.; WEN, F.; XIAO, R.; TIAN, O.; TANG, X. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 1222-1228, 2012.

DELAC, K.; GRGIC, M.; KOS, T. Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions. *In Proceedings of International Conference on Systems, Signals and Image Processing*, 95-98, 2006.

EVERINGHAM, M.; SIVIC, J.; ZISSERMAN, A. Hello! My name is... Buffy - Automatic naming of characters in TV video. *Proceedings of the 17th British Machine Vision Conference* (BMVC), 2006.

EVERITT, B. S.; LANDAU, S.; MORVEN, L. Cluster Analysis. *A Hodder Arnold Publication*, Vol. 4, 50-58, 2001.

FARFADE, S. S.; SABERIAN, M. D.; LI-JIA, LI. Multi-view Face Detection Using Deep Convolutional Neural Networks. *International Conference on Multimedia Retrieval* (ICMR), 2015.

FOUCHER, S.; GAGNON, L. Automatic Detection and Clustering of Actor Faces based on Spectral Clustering Techniques. *In Proceedings of the 4th Canadian Conference on Computer and Robot Vision*, 2007.

FUKUI, K.; YAMAGUCHI, O. The Kernel Orthogonal Mutual Subspace Method

and Its Application to 3D Object Recognition. *Lecture Notes in Computer Science* (ACCV), Vol. 4844, 467-476, 2007.

GAO, H.; EKENEL, H. K.; STIEFELHAGEN, R. Identifying Important People in Broadcast News Videos. *In Proceedings of the Conference on Machine Vision Applications* (IAPR), 127–136, 2011.

GEORGHIADES, A.S.; BELHUMEUR, P.; KRIEGMAN, D. From few to many: Illumination Cone Models For Face Recognition Under Variable Lighting And Pose. *In Proc. IEEE Transactions On Pattern Analysis And Machine Intelligence*, 2001.

GHAEMMAGHAMI, H.; DEAN, D.; VOGT, R.; SRIDHARAN, S. Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach. *In IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 4185–4188, 2012.

GHAEMMAGHAMI, H.; DEAN, D.; SRIDHARAN, S. Speaker attribution of australian broadcast news data. In 1st Workshop on Speech, Language and Audio in Multimedia (SLAM), 72–77, 2013.

GONZALEZ, R. C.; WOODS, R. E. Digital Image Processing. *Addison-Wesley Pub (Sd)*, Vol. 3, 2010.

GRAHAM, D.; ALLINSON, N. Characterizing virtual eigensignatures for general purpose face recognition. In Face Recognition: From Theory to Applications, Computer and Systems Sciences, 446–456, 1998.

GROSS, R.; SHI, J. The CMU motion of body (MoBo) database. *Robotics Institute*, Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-01-18, 2001.

HAMM, J.; LEE, D. D. Grassmann discriminant analysis: a unifying view on subspace-based learning. *In Proceedings of the International Conference on Machine Learning* (ICML), 376–383, 2008.

HAN, H.; SHAN, S.; CHEN, X.; GAO, W. A comparative study on illumination preprocessing in face recognition. *Pattern Recognition*, Vol. 46, 1691–1699, 2013.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, Vol. 143, 29–36, 1982.

HARANDI, M. T.; SANDERSON, C.; SHIRAZI, S.; LOVELL, B. C. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. *In Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition* (CVPR), 2705-2712, 2011.

HAR-PELED, S.; KUMAR, N. Approximate Nearest Neighbor Search for Low Dimensional Queries. *Computational Geometry, Data Structures and Algorithms*, 2010.

HU, Y.; MIAN, A. S.; OWENS, R. Sparse approximated nearest points for image set classification. *In Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition*, 121-128, 2011.

HUANG, G.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report, 07-49, October, 2007.

HUANG, P.; WANG, Y.; SHAO, M. A New Method for Multi-view Face Clustering in Video Sequence. *In Proceedings of the IEEE International Conference on Data Mining Workshops* (ICDMW), 869-873, 2008.

JAIN, A. K. Algorithms For Clustering Data. *Prentice Hall*, New Jersey, 1991.

JAIN, V.; LEARNED-MILLER, E. FDDB: A Benchmark for Face Detection in Unconstrained Settings. *Technical Report UM-CS-2010-009, Dept. of Computer Science*, University of Massachusetts, 2010.

JUAN, L.; GWON, O. A Comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing* (IJIP), Vol. 3, 143-152, 2009.

KALAL, Z.; MIKOLAJCZYK, K.; MATAS, J. Tracking-Learning-Detection. *IEEE Transactions on Pattern Anal. Mach. Intell*, 34 (7), 1409-1422, 2012.

KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. *In IEEE International Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 1867–1874, 2014.

KHAN, K; GERBEN, S.; GLANVILLE, J.; SOWDEN, A.; KLEIJNEN, J. Undertaking Systematic Review of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition), NHS Centre for Reviews and Dissemination, University of York, IBSN 1 900640 20 1, 2001.

KIM, M.; KUMAR, S.; PAVLOVIC, V.; ROWLEY, H. Face Tracking and Recognition with Visual Constraints in Real-World Videos. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2008.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. *Engineering*, no. EBSE 2007-001, 1–57, 2007.

LAGANIÈRE, R. OpenCV 2 Computer Vision Application Programming Cookbook, *Packt Publishing*, UK, 2011.

LEE, K. C.; HO, J.; YANG, M. H.; KRIEGMAN, D. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, Vol. 99, 303–331, 2005.

LEVINE, D. M., BERENSON, M. L., e STEPHAN, D. Estatística: Teoria e aplicações. *LTC Editora*, Rio de Janeiro, 2000.

LI, H.; LIN, Z.; SHEN, X.; BRANDT, J.; HUA, G. A Convolutional Neural Network Cascade for Face Detection. *Computer Vision and Pattern Recognition* (CVPR), 2015.

LIN, D.; KAPOOR, A.; HUA, G.; BAKER, S. Joint People, Event, and Location Recognition in Personal Photo Collections Using Cross-Domain Context. *In Proceedings of the 11th European Conference on Computer Vision*, 2010.

LOS ANGELES TIMES. YouTube by the number. Los Angeles Times, 31 de Maio, 2013. [Online] de http://articles.latimes.com/2011/may/31/business/la-fi-ct-youtube-kyncl-sidebar-20130531. Acesso em 05/09/2013.

LOWE, D. G. Object recognition from local scale-invariant features. *In Proceedings of the International Conference on Computer Vision*, 1150-1157, 1999.

MAFRA, S. N.; TRAVASSOS, G. H. Estudos primários e secundários apoiando a busca por evidência em Engenharia de Software. *Relatório Técnico*, RT-ES 687/06, 2006.

MARKUS, N.; FRLJAK, M.; PANDZIC, I. S.; AHLBERG, J.; FORCHHEIMER, R. A Method for Object Detection Based on Pixel Intensity Comparisons Organized in Decision Trees. *CoRR*, 2014.

MARTINKAUPPI, B.; SORIANO, M.; HUOVINEN, S.; LAAKSONEN, M. Face video database. *Proc. First European Conference on Color in Graphics, Imaging and Vision* (CGIV), Poitiers, France, 380-383, 2002.

MATTEUCCI, M. A Tutorial on Clustering Algorithms. Disponível em <http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html>. Acesso em 09/06/2011.

MIAN, A. Unsupervised Learning from Local Features for Video-based Face Recognition. *In Proceeding of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

MOURA, E.S.; GOMES, H. M.; DE CARVALHO, J. M. An Improved Face Verification Approach based on Speedup Robust Features and Pairwise Matching. *In Proceedings of the Conference on Graphics, Patterns and Images* (SIBGRAPI), 2013.

MUJA, M.; LOWE, D. G. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. *International Conference on Computer Vision Theory and Applications* (VISAPP), 2009.

NISHIYAMA, M.; YUASA, M.; SHIBATA, T.; WAKASUGI, T.; KAWAHARA T.; YAMAGUCHI, O. Recognizing Faces of Moving People by Hierarchical Image-Set Matching. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2007.

OJALA, T., PIETIKAINEN, M. e HARWOOD, D. A comparative-study of texture measures with classification based on feature distributions. *Pattern Recognition*, vol. 29, 1996.

OTTO, C.; WANG, D.; JAIN, A. K. Clustering Millions of Faces by Identity. *Computer Vision and Pattern Recognition* (CVPR), 2016.

PAI, M.; MCCULLOCH, M.; COLFORD, J. Systematic Review: A Road Map Version 2.2. *Systematic Reviews Group*, UC Berkeley, 2004.

PHILLIPS, P. J. ET AL. Overview of the multiple biometrics grand challenge. *In Proceedings of the International Conference of Advances Biometrics*, 705–714, 2009.

RAMANAN, D.; BAKER, S.; KAKADE, S. Leveraging archival video for

building face datasets. *In Proceedings of the IEEE 11th International Conference on Computer Vision* (ICCV), 1-8, 2007.

RIESENHUBER, M. e POGGIO, T. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2: 1019-1025, 1999.

SAMARIA, F. e HARTER, A. Parameterization of a stochastic model for human face identification. *In Proceeding of the IEEE Workshop on Applications of Computer Vision*, 1994.

SAMPAIO, R. F.; MANCINI, M. C. Estudos de Revisão Sistemática: Um guia para Síntese Criteriosa da Evidência Científica. *Revista Brasileira de Fisioterapia*. São Carlos, Vol. 11, 83-89, 2007.

SCHROEDER, S. YouTube in Numbers: 1 Month, 100 Million US Viewers, 6.3 Billion Videos. *Mashable Social Media*, 5 de Março, 2009. [Online] de http://mashable.com/2009/03/05/youtube-100-million/. Acesso em 17/11/2011.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2015.

SILVA, ALICE. Análise Classificatória. *Universidade Nova de Lisboa - Faculdade de Ciências e Tecnologia*, 2005. Disponível em <http://ferrari.dmat.fct.unl.pt/services/AnaliseDados/Cluster.pdf>. Acesso em 28/12/2011.

SIM, T.; BAKER, S.; BSAT, M. The CMU pose, illumination, and expression database. *IEEE Transactions Pattern Anal. Mach. Intell.*, Vol. 25, 1615–1618, 2003.

SONY, A.; AJITH, K.; THOMAS, K.; THOMAS, T.; OEEPA P. L. Video Summarization By Clustering Using Euclidean Distance*. In Proceedings of the 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies* (ICSCCN), 642-646, 2011.

SWAIN, M. J.; BALLARD, B. H. Color indexing. *International Journal of Computer Vision*, Vol. 7, 11-32, 1991.

TANG, Z.; YIFAN, Z.; ZECHAO, L.; HANQING L. Face clustering in videos with proportion prior. *In Proceedings of the 24th International Conference on Artificial Intelligence* (IJCAI), 2191-2197, 2015.

TAO, J.; TAN, Y-P. Face Clustering in Videos using Constraint Propagation. *In Proceedings of the IEEE International Symposium on Circuits and Systems* (ISCAS), 3246-3249, 2008.

TAPASWI, M.; PARKHI, O. M.; RAHTU, E.; SOMMERLADE, E.; STIEFELHAGEN, R.; ZISSERMAN, A. Total Cluster: A person agnostic clustering method for broadcast videos. *In Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing* (ICVGIP), 2014.

THAI, H.; TRUONG, N. Face Alignment Using Active Shape Model And Support Vector Machine. *In Proceedings of the International Journal of*

Biometrics and Bioinformatics, Vol. 4, 224-234, 2011.

THEODORIDIS, S.; KOUTROUMBAS, K. Pattern Recognition. Academic Press, 1999.

TURAGA, P.; VEERARAGHAVAN, A.; CHELLAPPA, R. Unsupervised View and Rate Invariant clustering of Video Sequences. *Journal Computer Vision and Image Understanding*, Vol. 113, 2009.

TUZEL, O.; PORIKLI, F.; MEER, P. Region covariance: A fast descriptor for detection and classification. *In Proceedings of the European Conference on Computer Vision* (ECCV), 589–600, 2006.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), Vol. 1, 511–518, 2001.

WANG, T.; SHI, P. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, Vol. 30, 1161–1165, 2009.

WANG, R.; SHAN, S.; CHEN, X.; GAO, W. Manifold-Manifold Distance with application to face recognition based on image set. *In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), 2008.

WANG, L.; YAN, H.; LV, K.; PAN, C. Visual Tracking Via Kernel Sparse Representation With Multikernel Fusion. *In IEEE Transactions on Circuits and Systems for Video Technology*, 1132-1141, 2014.

WEBSITE MONITORING BLOG. YouTube Facts & Figures (history & statistics). *Website Monitoring Blog*, 17 de Maio, 2010. [Online] de http://www.website-monitoring.com/blog/2010/05/17/youtube-facts-and-figures-history-statistics/. Acesso em 17/11/2011.

WECHSLER, H. Reliable Face Recognition Methods: System Design, Implementation and Evaluation (International Series on Biometrics). *Springer-Verlag*, New York, Secaucus, NJ, USA, 2006.

WEINBERGER, K. Q.; BLITZER, J.; SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press, 2006.

WOLF, L.; HASSNER, T.; MAOZ, I. Face recognition in unconstrained videos with matched background similarity. *In Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition*, 529-534, 2011.

YAMAMOTO, K.; YAMAGUCHI, O.; AOKI, H. Fast face clustering based on shot similarity for browsing video. *In Proceedings of the Progress in Informatics, Special issue: 3D image and video technology*, Vol. 7, 53–62, 2010.

YOUTUBE ADVERTISE. Youtube Advertise. [Online] de http://www.youtube.com/advertise. Acesso em 23/09/2014.

YUSOFF, Y.; CHRISTMAS, W.; KITTLER, J. A Study on Automatic Shot Change Detection. *Multimedia Applications, Services and Techniques*,

1998.

ZHANG, Y-F.; XU, C.; LU, H.; HUANG, Y-M. Character Identification in Feature-Length Films Using Global Face-Name Matching. *In Proceedings of the IEEE Transactions on Multimedia*, Vol. 11, 2009.

ZHAO, W.; CHELLAPPA, R.; PHILLIPS, P.; ROSENFELD, A. Face recognition: A literature survey. *ACM Computing Surveys*, Vol. 35, 339-458, 2003.

ZHONG, W.; LU, H.; YANG, M. Robust object tracking via sparsity based collaborative model. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), 1838–1845, 2012.

ZHOU, C.; ZHANG, C.; FU, H.; WANG, R.; CAO, X. Multi-cue Augmented Face Clustering. *In Proceedings of the 23rd ACM international conference on Multimedia* (MM), 1095-1098, 2015.

Apêndice A

Métricas de Avaliação de Agrupamento

Este apêndice contém uma descrição sucinta das métricas de avaliação de agrupamento utilizados nesta tese (com exceção das métricas *Average Purity* e *Average Coverage*, que foram definidas na Seção 4.3.3), a saber: (i) *Rand Index* (RI); (ii) *Adjusted Rand Index* (ARI); (iii) *Precision* (P); (iv) *Recall* (R); (v) *F-Mesure* (F); (vi) *Jaccard Index* (JI); e (vii) *Folkes and Mallows Index* (FMI).

Uma das principais dificuldades em problemas de classificação consiste na correta avaliação do desempenho do classificador. Isto geralmente é feito a partir da adoção de uma medida de desempenho comum, e.g., erro médio quadrático (MSE), área em porcentagem sob a curva ROC (*Receiver Operating Characteristic*), dentre outras. Todas estas métricas permitem comparar o resultado rotulado pelo algoritmo de classificação supervisionada com rótulos previamente conhecidos (*ground truth*).

Diferentemente, para algoritmos de agrupamento não supervisionado torna-se útil a introdução de uma medida que permita avaliar o desempenho do algoritmo em relação à qualidade da divisão dos dados de entrada em classes ou grupos diferentes, focando na relação entre os elementos de cada classe e não em rótulos previamente fornecidos. Esta é a razão principal da utilização de métricas de validação de agrupamentos em problemas de classificação não supervisionada. As métricas descritas a seguir foram escolhidas conforme resultado da análise de trabalhos revisados anteriormente.

Rand Index (RI) é uma métrica da similaridade entre dois agrupamentos de dados. Dados um conjunto de n elementos $S = \{o_1,...,o_t\}$ e duas partições de S a comparar, $X = \{x_1,...,x_r\}$ e $Y = \{y_1,...,y_n\}$ (em que

x_i e y_i denotam agrupamentos de elementos de S), consideram-se as seguintes propriedades:

(i) *a*, número de pares de elementos em *S* que estão no mesmo conjunto em *X* e no mesmo conjunto em *Y*;

(ii) *b*, número de pares de elementos em *S* que estão no mesmo conjunto em *X* e em conjuntos distintos em *Y*;

(iii) *c*, número de pares de elementos em *S* que estão em conjuntos distintos em *X* e no mesmo conjunto em *Y*; e

(iv) *d*, número de pares de elementos em *S* que estão em conjuntos distintos em *X* e em conjuntos distintos em *Y*.

A métrica *Rand Index* (RAND, 1971) é definida como sendo:

$$RI = \frac{\boldsymbol{a}+\boldsymbol{d}}{\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}+\boldsymbol{d}} = \frac{\boldsymbol{a}+\boldsymbol{d}}{\binom{\boldsymbol{n}}{2}} \qquad \textbf{(B.1)}$$

Intuitivamente, pode-se considerar a soma *a* + *d* como sendo o número de concordâncias entre *X* e *Y* e *b* + *c* como sendo o número de discordâncias entre *X* e *Y*. O índice Rand varia no intervalo [0,1], com 0 indicando que os dois conjuntos de dados não concordam com qualquer par de pontos e 1 indicando que os agrupamentos de dados são exatamente iguais.

De fato, é desejável que o índice de similaridade entre as duas partições aleatórias assuma valor próximo de zero. O problema com o índice de Rand é que seus valores esperados para pares de partições aleatórias nem sempre são discriminantes, ou seja, geralmente são gerados valores próximos. Hubert e Arabie (1985), ao tomar a distribuição hipergeométrica generalizada como o modelo de aleatoriedade, encontraram o valor esperado para *a* + *d* e sugeriram uma correção na métrica *Rand*, no sentido de aumentar o poder de diferenciação entre duas partições aleatórias.

Adjusted Rand Index (ARI) é um melhoramento da métrica *Rand* e se tornou um dos índices de avaliação de agrupamentos mais utilizados e recomendados para a medição de concordância entre duas partições com

diferentes números de grupos. ARI (SANTOS e EMBRECHTS, 2009) é formulado como sendo:

$$ARI = \frac{Index - Expected\_Index}{Max\_Index - Expected\_Index} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \quad \textbf{(B.2)}$$

De maneira similar à métrica RI, a métrica ARI assume valores de 0 (agrupamentos totalmente diferentes) a 1 (agrupamentos idênticos).

Precision-Recall (PR), ou precisão e revocação, são duas métricas utilizadas para avaliar a corretude de um algoritmo de reconhecimento de padrões. Elas podem ser vistas como versões estendidas da acurácia, uma métrica simples que calcula a fração de instâncias na qual o resultado correto é retornado (DAVIS e GOADRICH, 2006). Precisão mede a habilidade de o sistema retornar apenas resultados relevantes (OLSEN e DELEN, 2008), podendo ser definida como:

$$P = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + Falso\ Positivo} \quad \textbf{(B.3)}$$

Cobertura mede a habilidade de o sistema retornar todos os resultados relevantes (OLSEN e DELEN, 2008) e pode ser formulada como:

$$R = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + Falso\ Negativo} \quad \textbf{(B.4)}$$

F-Mesure (F) é a média harmônica de precisão e cobertura, também conhecida como *F1 Score*. A métrica F (KANDEFER e SHAPIRO, 2009), pode ser definida como:

$$F = 2.\frac{P.R}{P+R} \quad \textbf{(B.5)}$$

Similarmente às métricas *Rand*, assume valores no intervalo [0,1].

Jaccard Index (JI), também conhecido como o coeficiente de similaridade de *Jaccard*, é uma métrica utilizada para comparação de

similaridade e diversidade entre conjuntos de amostras, e é definido como o tamanho da interseção dividido pelo tamanho da união do conjunto de amostras. O índice de *Jaccard* (IVCHENKO e HONOV, 1998), pode ser expresso por:

$$J = \frac{a}{a+b+c} \quad \textbf{(B.6)}$$

Da mesma forma que as métricas *Rand* e *F*, o índice de *Jaccard* varia no intervalo [0,1].

Folkes and Mallows Index (FM) é um método de avaliação externa utilizado para determinar a similaridade entre os dois agrupamentos. Esta métrica de similaridade pode mensurada entre dois agrupamentos hierárquicos ou entre um agrupamento e uma classificação de referência. Quanto maior o valor do índice Fowlkes-Mallows, mais próximos são os agrupamentos (ou *clusters*). O índice de *Folkes and Mallows* (HALKIDI, BATISTAKIS e VAZIRGIANNIS, 2011), pode ser definido como sendo:

$$FM = \frac{a}{\sqrt{m_1 m_2}} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \quad \textbf{(B.7)}$$

sendo $m_1 = (a+b)$ e $m_2 = (a+c)$. Da mesma forma que as métricas *Rand*, *F* e *Jaccard*, o índice de *Folkes* e *Mallows* varia no intervalo [0,1].

Apêndice B

Protocolo de Estudo

Neste apêndice é apresentado o protocolo da revisão bibliográfica realizada. O protocolo de estudo deve ser elaborado a *priori* e tem como objetivo registrar de forma clara e transparente todo o processo que envolve a realização da revisão bibliográfica, a definição dos objetivos do estudo, a organização dos procedimentos a serem seguidos pelos executores da revisão, bem como definir as análises que serão realizadas.

O presente protocolo de pesquisa foi definido de acordo com a seguinte metodologia: (i) identificação do objetivo principal e objetivos secundários; (ii) definição das questões de pesquisa da RS; (iii) seleção das bases de dados a serem pesquisadas; (iv) elaboração da estratégia (*strings*) de busca; (v) definição de critérios de elegibilidade (inclusão e exclusão); (vi) descrição do processo de triagem e seleção de publicações; e (vii) execução do processo de extração e análise de dados.

B.1. Objetivos

Uma RS objetiva identificar, avaliar e interpretar toda pesquisa disponível e relevante sobre uma questão de pesquisa, um tópico ou um fenômeno de interesse (MAFRA e TRAVASSOS, 2006). A conduta de uma RS apresenta uma avaliação justa do tópico de pesquisa, à medida que utiliza uma metodologia de revisão rigorosa, confiável e passível de auditagem (KITCHENHAM e CHARTERS, 2007). A RS permite, também, a reutilização de seu conteúdo, parte e/ou integralmente, por outro usuário, mantendo assim, a estrutura e informações de um estudo revisado, consistente e reprodutível.

Nesse sentido, o objetivo principal do presente apêndice consiste na

realização de uma RS da literatura da área, a fim de organizar e sintetizar informações, apresentar um panorama geral e uma nova perspectiva acerca dos estudos existentes, identificar eventuais limitações e problemas em aberto partir de uma análise crítica e comparativa de estudos relacionados, assim como potenciais soluções existentes no estado da arte sobre Agrupamento de Faces em Vídeos.

Além da condução da RS, destacam-se os seguintes objetivos específicos a serem atingidos: (i) identificar os trabalhos mais recentes e relevantes sobre Agrupamento de Faces em Vídeos; (ii) analisar similaridades e pontos de diferenciação entre as abordagens, métodos e técnicas propostos nos trabalhos relacionados; (iii) identificar melhores práticas recomendadas por essas diferentes abordagens; (iv) contribuir para comunidade científica com a identificação de desafios e lacunas de pesquisa que necessitem de maior exploração; e (v) propor uma nova abordagem para o problema em questão, objetivando a obtenção do melhor desempenho em relação ao estado da arte.

B.2. Questões de Pesquisa

A definição das questões de pesquisa é a parte mais importante de uma Revisão Sistemática, uma vez que orienta todo o processo da metodologia da revisão (elaboração do projeto, identificação e seleção dos estudos, extração dos dados, avaliação da qualidade, análise, apresentação e interpretação dos resultados), bem como critério de avalição da relevância da pesquisa (KITCHENHAM e CHARTERS, 2007). Diante dos objetivos propostos anteriormente, a execução desta RS visa dar subsídios para responder a questão principal (QP) deste estudo:

- QP: Realizar uma RS sobre o tema de Agrupamento de Faces em Vídeos, a fim de identificar quais são os aspectos (técnicas, métodos e componentes) a serem considerados na concepção de um sistema capaz de obter melhor desempenho em relação ao estado da arte.

Com base na questão principal de pesquisa, foram levantadas

questões secundárias (QS), conforme descrito no Quadro B.1.

Quadro B.1 – Questões secundárias de pesquisa.

Questão Secundária	Motivações
QS1: Quais são as técnicas/métodos de agrupamento adotados em sistemas de agrupamento de faces em vídeos digitais?	Identificar e avaliar as melhores técnicas/métodos de agrupamento que irão compor o sistema a ser desenvolvido.
QS2: Quais são as características extraídas para composição de identificação pessoal?	Identificar características mais frequentes em sistemas de agrupamento de faces em vídeos digitais, objetivando a caracterização única de cada face.
QS3: Quais são as métricas utilizadas para medir com precisão a similaridade entre faces?	Definir as principais métricas que avaliam numericamente o grau de semelhança entre faces objetivando a extensão da métrica existente ou a elaboração de uma nova métrica inspirada nas demais.
QS4: Quais são as métricas utilizadas para avalição dos grupos gerados por sistemas de agrupamento de faces em vídeos digitais?	Definir as métricas que serão utilizadas nos experimentos de comparação, tais como: (i) Rand Index; (ii) Adjusted Rand Index; (iii) Precision; (iv) Recall; (v) F-Mesure; (vi) Jaccard Index; e (vii) Folkes and Mallows Index.
QS5: Quais são as bases de dados utilizadas para testes de sistemas de agrupamento de faces em vídeos digitais?	Definir quais bases de dados pública mais comumente utilizadas para posterior comparação de resultados.

B.3. Seleção dos Engenhos de Busca

A pesquisa foi realizada por meio de buscas manuais e eletrônicas de publicações relevantes ao escopo deste estudo. Foram considerados trabalhos publicados em conferências, simpósios, revistas e capítulos de livros no período de 2007 a 2016. As bibliotecas digitais foram acessadas em seus *websites* oficiais, realizando pesquisas por meio dos seus respectivos engenhos de busca avançada (*advanced search*). A mesma lógica da expressão de busca foi utilizada em todas as bibliotecas digitais abordadas, porém, algumas adaptações foram necessárias para ajustar a expressão ao padrão adotado pelo engenho de busca correspondente.

As bibliotecas digitais que foram consultadas nesta RS apresentam as seguintes características como critério de inclusão nesta pesquisa:

- O engenho de busca deve permitir a utilização de expressões

lógicas (OR, AND e NOT) e o agrupamento de comandos por meio de parênteses ou similares;

- O engenho de busca deve permitir a realização de consultas em meta-dados das publicações, e.g., título, resumo, palavras-chave, dentre outros;
- O engenho de busca deve possuir algum recurso para a exportação dos dados e resumos de todas as publicações selecionadas pela expressão de busca;
- O engenho de busca deve possuir abrangência em revistas e conferências mais importantes e com os maiores impactos na comunidade científica.

As bases de dados bibliográficas digitais encontradas que atenderam aos critérios seleção são as listadas a seguir:

- *IEEEXplore* (http://ieeexplore.ieee.org);
- *Science Direct* (http://www.sciencedirect.com);
- *Scopus* (http://www.scopus.com);
- *Web of Science* (http://webofknowledge.com);
- *Compendex* (http://www.engineeringvillage2.org);
- *ProQuest* (http://search.proquest.com).

Após a identificação das bases de dados bibliográficas digitais, foram realizadas as coletas dos estudos pela expressão de busca, utilizando os engenhos de busca das bases de dados identificadas, as publicações obtidas foram analisadas para determinar suas relevâncias para a RS.

B.4. Identificação das Palavras-Chave de Busca

A expressão de busca é um conjunto composto por palavras-chave (*strings*), as quais são utilizadas para realizar pesquisas por publicações científicas disponíveis em bibliotecas digitais. A partir da definição das questões de pesquisa foi identificado um conjunto de palavras-chave, assim como expressões que definem a linha de pesquisa com o objetivo de ampliar os

resultados retornados pela expressão de busca a ser elaborada, a saber: (i) *face clustering*; e (ii) *vídeo*. Adicionalmente, identificou-se um conjunto de sinônimos para as palavras-chave, assim, a *string* de busca foi construída a partir de uma combinação das palavras-chave e de seus sinônimos, conforme representado no Quadro B.2.

A expressão de busca desta RS é formada pelo agrupamento dos termos-chaves dentro de cada grupo de palavras-chave. Cada grupo contém termos que são sinônimos, termos que têm significado semântico similar ou relacionado dentro do domínio. Os termos escolhidos devem ser intimamente relacionados com as questões de pesquisas. Com a definição dos termos a serem utilizados, os resultados da busca foram combinados utilizando os operadores booleanos, especificamente, o "OR" e o "AND".

Quadro B.2 – Palavras-chave e termos de busca.

Termos \ Grupos	G1	G2	G3
T1	*face clustering*	*video*	*evaluation*
T2	*face matching*	*video retrieval*	*performance*
T3	*face tracking*	*video summary*	*metric*
T4	-	*video search*	*measure*
T5	-	*movies*	*comparison*
T6	-	*real-time*	*benchmark*
T7	-	*tracks*	*experimental*

De posse dos termos de busca e grupos de sinônimos presentes no Quadro B.2, a expressão de busca adotada nesta RS é definida como segue:

(face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental)

A seguir, são apresentadas as expressões de busca para cada um dos engenhos de busca considerados nesta RS:

- *IEEEXplore:*

((p_Title:"face clustering" OR p_Title:"face matching" OR p_Title:"face tracking" OR p_Abstract:"face clustering" OR p_Abstract:"face matching" OR

p_Abstract:"face tracking") AND (p_Title:"video" OR p_Title:"video retrieval" OR p_Title:"video summary" OR p_Title:"video search" OR p_Title:"movies" OR p_Title:"real-time" OR p_Title:"tracks" OR p_Abstract:"video" OR p_Abstract:"video retrieval" OR p_Abstract:"video summary" OR p_Abstract:"video search" OR p_Abstract:"movies" OR p_Abstract:"real-time" OR p_Abstract:"tracks") AND (p_Title:"evaluation" OR p_Title:"performance" OR p_Title:"metric" OR p_Title:"measure" OR p_Title:"comparison" OR p_Title:"benchmark" OR p_Title:"experimental" OR p_Abstract:"evaluation" OR p_Abstract:"performance" OR p_Abstract:"metric" OR p_Abstract:"measure" OR p_Abstract:"comparison" OR p_Abstract:"benchmark" OR p_Abstract:"experimental"))

- *ScienceDirect:*

TITLE-ABSTR-KEY((face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental))

- *Scopus:*

(TITLE-ABS-KEY(face clustering) OR TITLE-ABS-KEY(face matching) OR TITLE-ABS-KEY(face tracking)) AND (TITLE-ABS-KEY(video) OR TITLE-ABS-KEY(video retrieval) OR TITLE-ABS-KEY(video summary) OR TITLE-ABS-KEY(video search) OR TITLE-ABS-KEY(movies) OR TITLE-ABS-KEY(real-time) OR TITLE-ABS-KEY(tracks)) AND (TITLE-ABS-KEY(evaluation) OR TITLE-ABS-KEY(performance) OR TITLE-ABS-KEY(metric) OR TITLE-ABS-KEY(measure) OR TITLE-ABS-KEY(comparison) OR TITLE-ABS-KEY(benchmark) OR TITLE-ABS-KEY(experimental))

- *Web of Science:*

(TI=((face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental)) OR TS =((face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video

summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental))

- *Compendex:*

(face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental)

- *ProQuest:*

ti((face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental)) OR ab((face clustering OR face matching OR face tracking) AND (video OR video retrieval OR video summary OR video search OR movies OR real-time OR tracks) AND (evaluation OR performance OR metric OR measure OR comparison OR benchmark OR experimental))

B.5. Critérios de Elegibilidade (Inclusão e Exclusão)

As publicações armazenadas nas bibliotecas digitais, previamente selecionadas com o auxílio da expressão de busca, foram catalogadas na base de dados do sistema *StArt*[8], para auxiliar os pesquisadores na detecção de publicações duplicadas, nos processos de filtragens e em análises futuras.

Como a seleção prévia das publicações por intermédio da expressão de busca não garantia que todo o material coletado esteja alinhado ao escopo desta RS, fez-se necessária à aplicação de filtros para excluir publicações fora do contexto. As publicações foram classificadas por meio de critérios de inclusão e exclusão a partir da leitura de meta-dados de cada

[8] StArt, disponível em: *http://lapes.dc.ufscar.br/tools/start_tool*

publicação selecionada pela expressão de busca, i.e., o título, o resumo e as palavras-chave.

Um sumário dos critérios de inclusão e de exclusão empregados nesta Revisão Sistemática é apresentado no Quadro B.3. Foram consideradas nesta RS as publicações previamente selecionadas que apresentaram todos os critérios de inclusão e não satisfizeram qualquer um dos critérios de exclusão.

Quadro B.3 – Critérios de inclusão e exclusão.

Critério de Inclusão (CI)	Critério de Exclusão (CE)
CI1: A publicação descreve uma pesquisa relacionada ao problema objeto de estudo.	**CE1:** A publicação contém palavras-chave presentes exclusivamente nas seções de biografia dos autores, agradecimentos, referências bibliográficas ou anexos.
CI2: A publicação apresenta uma solução para o tema abordado nesta RS.	**CE2:** A publicação não possui relevância com relação ao tema abordado nesta revisão.
CI3: A publicação apresenta um procedimento de teste para avaliação da solução proposta.	**CE3: A** publicação não satisfaz qualquer uma das questões de pesquisa.
CI4: A publicação apresenta resultados empíricos claros e completos que podem ser comparados.	**CE4:** A publicação não apresenta uma avaliação experimental clara e bem definida.
CI5: A publicação responde parcialmente ou integralmente as questões de pesquisa.	**CE5: A** publicação apresenta resultados superficiais e imprecisos que inviabilizam sua comparação.
CI6: A publicação utiliza uma base de dados pública.	**CE6:** A publicação utiliza apenas base de dados proprietária.

B.6. Processo de Triagem e Seleção de Publicações

Após a definição das questões de pesquisa, da estratégia de busca e dos critérios de inclusão e exclusão, as publicações selecionadas pela expressão de busca foram analisadas com base em sua relevância e em função de critérios de qualidade. O processo de triagem e seleção das publicações seguiu a metodologia de Kitchenham e Charters (2007), de acordo com as etapas a seguir:

- **Etapa 1:** Identificar publicações relevantes nas bibliotecas digitais selecionadas;
- **Etapa 2:** Filtrar as duplicadas, usando a ferramenta StArt;

- **Etapa 3:** Filtrar com base nos seus títulos;
- **Etapa 4:** Filtrar com base nos seus resumos (*abstracts*);
- **Etapa 5:** Filtrar com base na introdução, análise experimental e conclusão;
- **Etapa 6:** Obter os estudos primários e analisar criticamente os trabalhos.

Na Figura B.1, ilustra-se o processo de triagem e seleção de publicações adotado nesta RS contendo informações do número de publicações identificadas e filtradas a cada etapa e os critérios de exclusão aplicados.

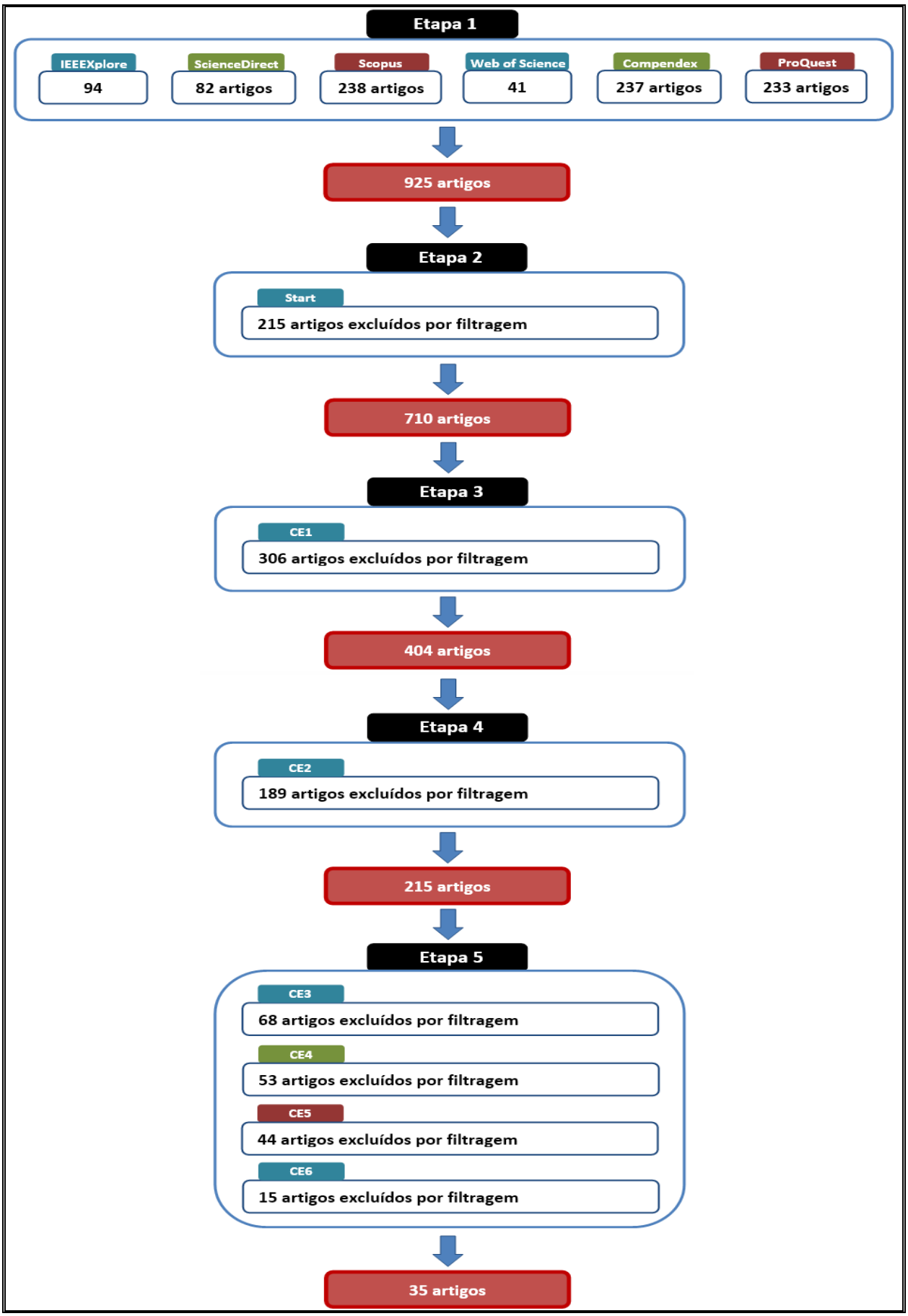
Após as cinco etapas do processo de filtragem, descritas anteriormente, foi aplicada a etapa final de seleção que consistiu na análise crítica após leitura completa das publicações remanescentes. Para as bibliotecas digitais *Scopus* e *Compendex* foram consideradas apenas publicações em periódicos devido ao grande volume de dados retornados.

Na Etapa 1, realizou-se a busca avançada nas bases de bibliografias digitais listadas na Seção B.3, considerando as respectivas expressões de busca para cada engenho, conforme descrito na Seção B.4. Foram armazenadas todas as informações relevantes (título, autores, fonte, ano, *abstract* e palavras-chave) dos 925 artigos identificados na Etapa 1 na ferramenta *StArt*. Na Etapa 2, foram removidas as publicações duplicadas (mesmo título) ou semelhantes (evolução do mesmo artigo), neste último, foi aceita versão mais atualizada do estudo, resultando em 710 artigos únicos.

Na Etapa 3, foi aplicado o Critério de Exclusão 1 (CE1), a partir do qual 306 publicações foram excluídas, resultado em 404 artigos para posterior análise. Na Etapa 4, foi aplicado o Critério de Exclusão 2 (CE2), a partir do qual 189 publicações foram excluídas, resultado em 215 artigos a serem filtrados pela etapa seguinte. Na Etapa 5, foram aplicados os critérios de exclusão CE3, CE4 e CE5, a partir do qual 180 publicações foram excluídas, resultando em 35 publicações relacionadas com o objeto de

estudo que foram submetidos a uma análise qualitativa (Etapa 6).

Figura B.1 – Processo de triagem e seleção de publicações.



B.7. Avaliação de Qualidade

Após o processo de triagem de publicações, foram utilizados critérios de avaliação de qualidade para avaliar os estudos selecionados com o objetivo de guiar a interpretação dos estudos de maneira a mensurar a credibilidade dos mesmos quanto à qualidade e à relevância de seus resultados.

Os critérios de avaliação de qualidade foram utilizados para obter uma pontuação final para cada estudo resultante da etapa anterior. O procedimento de pontuação das questões de avaliação foi de S (sim) = 1, P (parcialmente) = 0.5 e N (não) = 0, conforme proposto por Kitchenham e Charters (2007). Os artigos com pontuação 0 foram retirados do estudo.

Os critérios utilizados nesta RS foram:

- Foi utilizada alguma metodologia para dar suporte à abordagem apresentada?
- Foi abordado algum dispositivo, instrumento, ferramentas para auxiliar a abordagem apresentada?
- Foi utilizada alguma métrica de desempenho?
- As métricas de desempenho utilizados no estudo são explicadas e justificadas?
- As observações / resultados sustentam as conclusões?

Com a execução do processo de avaliação de qualidade segundo os critérios listados 30 publicações (vide Figura B.1) foram selecionadas como altamente relevantes com a presente pesquisa.

Apêndice C

Fundamentação Teórica

Neste apêndice é apresentada uma descrição sucinta de técnicas e métodos das áreas de Processamento Digital de Imagens (PDI) e Visão Computacional (VC) utilizados nesta tese.

C.1. Filtragem Homomórfica

Uma imagem é formada a partir da luz refletida pelos componentes da cena retratada, podendo ser caracterizada por dois parâmetros: (i) a quantidade de iluminação que incide na cena; e (ii) a quantidade de iluminação refletida pelos componentes da cena. Desta forma, uma imagem *f(x,y)* pode ser expressa como o produto dos componentes de iluminação, *i(x,y)*, e refletância, *r(x,y)* (GONZALEZ e WOODS, 2010), ou seja:

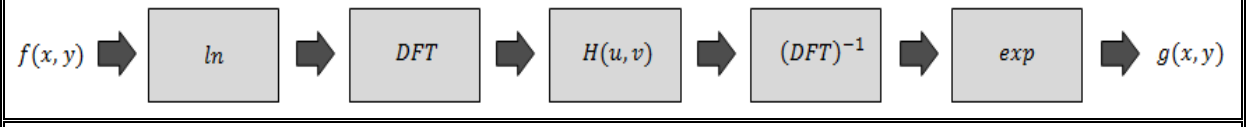
$$f(x,y) = i(x,y)r(x,y) \quad \text{(C.1)}$$

Este modelo de formação de imagens é conhecido como modelo de iluminação-refletância e tem sido usado para o melhoramento da qualidade de imagens adquiridas sob condições de iluminação não adequadas para a análise automática (DELAC, GRGIC e KOS, 2006).

O objetivo do módulo de compensação de iluminação é reduzir o componente de iluminação de uma imagem de face, de forma que a imagem final se apresente como uma aproximação da refletância da face, a qual é independente das condições de iluminação.

A filtragem homomórfica foi o procedimento utilizado para separar estes componentes, conforme apresentado no diagrama da Figura C.1. O processo de filtragem homomórfica permite manipular sinais combinados por intermédio de operações não-lineares.

Figura C.1 – Resumo dos passos da filtragem homomórfica.



Fonte: Adaptado de Gonzalez e Woods, 2010.

O algoritmo consiste em transformar o problema da combinação não-linear em um problema de combinação linear, utilizando o princípio da convolução por separabilidade e operações matemáticas apropriadas, em função de um filtro homomórfico *H(u,v)* (GONZALEZ e WOODS, 2010).

O componente de iluminação de uma imagem geralmente é caracterizado por variações espaciais suaves, enquanto o componente de refletância tende a variar abruptamente, particularmente nas junções de diferentes superfícies. Essas características levam a associar as componentes espectrais de baixa frequência de uma imagem à iluminação e aquelas de alta frequência à refletância.

Assim, a função da filtragem homomórfica é atenuar o componente de iluminação de uma imagem e ressaltar a refletância. O resultado da aplicação da filtragem homomórfica, presente na segunda etapa da abordagem proposta, é ilustrado na Figura C.2B.

Figura C.2 – Exemplo de compensação de iluminação: (A) Imagem normalizada; e (B) Imagem após filtragem homomórfica.

(A) **(B)**

C.2. Equalização de Histograma

A equalização de histograma é uma técnica a partir da qual se procura

redistribuir os valores de tons de cinza dos pixels em uma imagem, de modo a obter uma imagem cujo histograma se aproxime de uma distribuição uniforme, na qual o número (percentual) de pixels de qualquer nível de cinza é, idealmente, o mesmo (GONZALEZ e WOODS, 2010).

Para uma imagem *I*, contendo *N* pixels com um histograma *h(i)*, o histograma normalizado (GONZALEZ e WOODS, 2010), pode ser definido como:

$$\boldsymbol{h_{norm}(i)} = \frac{1}{\boldsymbol{n}}\boldsymbol{h(i)};\ \boldsymbol{i} = 0,\ldots,\boldsymbol{k}-1 \qquad \textbf{(C.2)}$$

O histograma normalizado ($h_{norm(i)}$) corresponde à densidade de probabilidade de valores de cinza dos pixels da imagem. O histograma acumulado é definido em termos do histograma normalizado como:

$$\boldsymbol{h_{acum}(r)} = \sum_{i=0}^{r}\boldsymbol{h_{norm}(i)};\ \boldsymbol{r} = 0,\ldots,\boldsymbol{k}-1 \qquad \textbf{(C.3)}$$

em que a função h_{acum} é uma função de distribuição de probabilidade acumulada. Portanto, trata-se uma função não decrescente com $h_{acum(k-1)} = 1$. A partir da Equação (C.3), o histograma normalizado é expresso como:

$$\boldsymbol{h_{norm}(q)} = \boldsymbol{h_{acum}(q)} - \boldsymbol{h_{acum}(q-1)};\ \ \boldsymbol{q} = 0,\ldots,\boldsymbol{k}-1 \qquad \textbf{(C.4)}$$

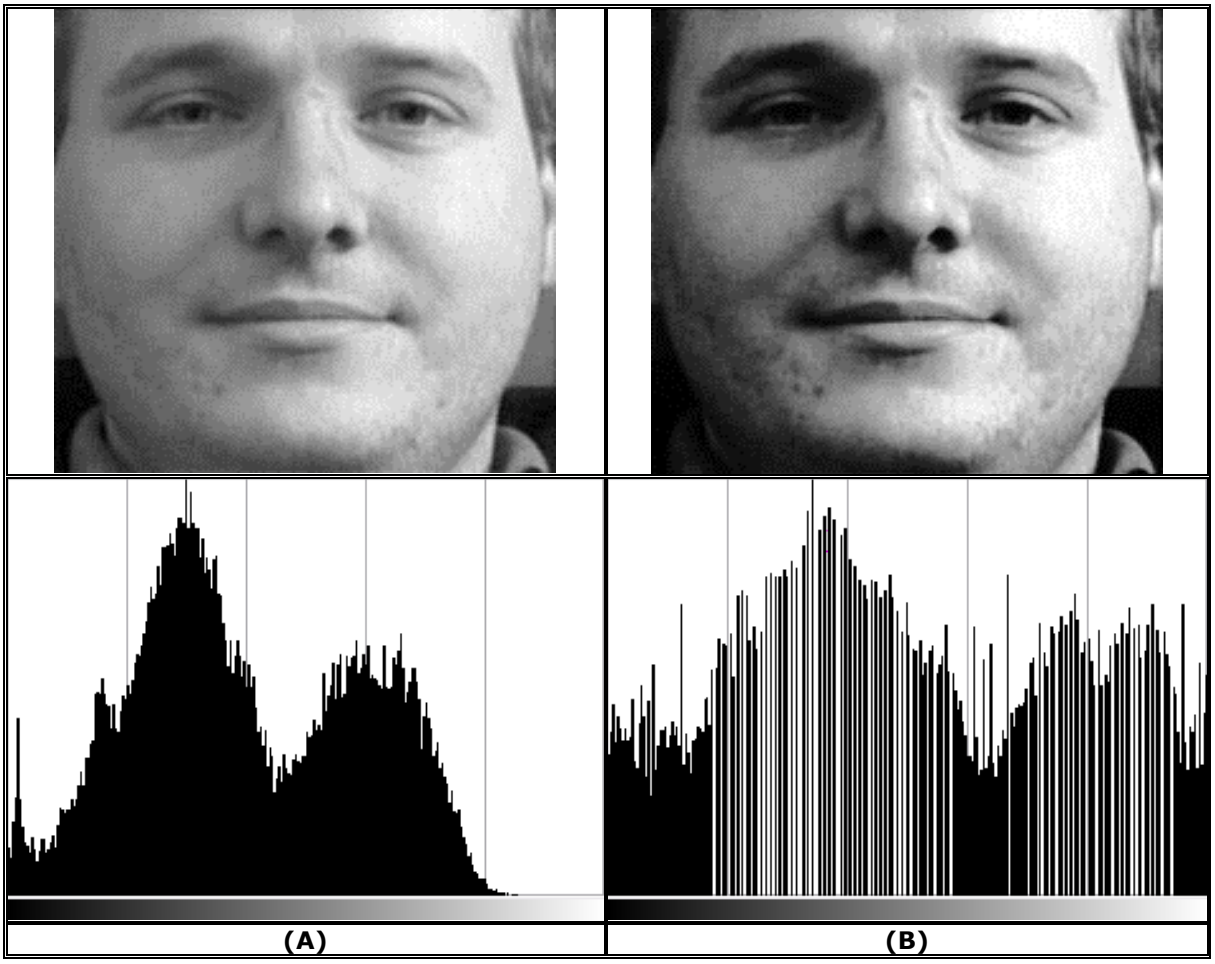
Tratando-se h_{norm} e h_{acum} como funções de uma variável contínua *x*, estes vetores podem ser vistos como uma função densidade de probabilidade (FDP) e uma função de distribuição acumulada (FDA); assim, a diferenciação *d*/*dx* corresponde ao relacionamento h_{norm} = d$h_{acum(x)}/dx$. Portanto, a imagem produzida pela operação de equalização pode ser descrita por:

$$\boldsymbol{E(x,y)} = \boldsymbol{eq(I(x,y))} = \boldsymbol{h_{acum}(I(x,y))} \qquad \textbf{(C.5)}$$

em que a imagem *E* tem um histograma com distribuição que se aproxima da distribuição uniforme. O resultado da aplicação da equalização de

histograma, presente na segunda etapa da abordagem proposta, é ilustrado na Figura C.3. Pode-se observar que na imagem resultante do processamento, alguns detalhes das características faciais são realçados.

Figura C.3 – Exemplo de equalização de histograma: (A) Imagem após filtragem homomórfica; e (B) Imagem após equalização de histograma.



C.3. *Fast Approximate Nearest Neighbors* – FANN

O problema do vizinho mais próximo (*Nearest Neighbors*) pode ser formulado como segue. Dado um conjunto P de n pontos em um espaço métrico X, pré-processar P, tal que, para um ponto de consulta $q \in X$, pode-se encontrar (rapidamente) o ponto $n_q \in P$ mais próximo de q. A busca de vizinhos mais próximos é uma tarefa fundamental utilizada em vários domínios, incluindo aprendizagem de máquina, agrupamento de objetos, recuperação de documentos, bancos de dados, estatísticas, dentre outros (HAR-PELED e KUMAR, 2010).

Em algumas aplicações, pode ser aceitável a recuperação de um "bom palpite" do vizinho mais próximo. Nessas situações, pode-se usar um

algoritmo que não garanta encontrar o vizinho mais próximo real em todos os casos, em troca de maior eficiência em termos de velocidade ou memória.

Tal algoritmo vai encontrar o vizinho mais próximo, na maioria dos casos, mas isso depende muito do conjunto de dados que está sendo consultado. Um exemplo é o FANN, algoritmo que aplica busca de prioridade em árvores *K-Means* hierárquica (*Hierarchical K-Means Tree*), proposto por Muja e Loew (2009).

A árvore *K-Means* hierárquica é construída dividindo-se os pontos de dados de cada nível em *K* regiões distintas usando um agrupamento *K-Means* e, em seguida, aplicando-se o mesmo método recursivamente para os pontos em cada região.

A recursão é encerrada quando o número de pontos em uma região for menor do que *K*. O algoritmo FANN explora a árvore *K-Means* hierárquica da forma *best-bin-first* (com base em árvores *kd*), ou seja, retorna o vizinho mais próximo para uma grande fração de consultas e um vizinho muito próximo para os demais casos (HAR-PELED e KUMAR, 2010).

Nessa pesquisa, o algoritmo FANN realiza inicialmente um único percurso pela árvore, adicionando uma fila de prioridades a todos os ramos inexplorados em cada nó ao longo do caminho. Em seguida, extrai da fila de prioridades o ramo que possui o centro mais próximo ao ponto de consulta e reinicia a travessia da árvore a partir daquele ramo.

Em cada passagem, o algoritmo continua adicionando à fila de prioridades os ramos inexplorados ao longo do caminho. O grau de aproximação é especificado da mesma forma como é definido para árvores *kd* randomizadas, parando a pesquisa logo após um número predeterminado de nós folha (pontos de dados) serem examinados (MUJA e LOEW, 2009).

C.4. Rastreador SURF

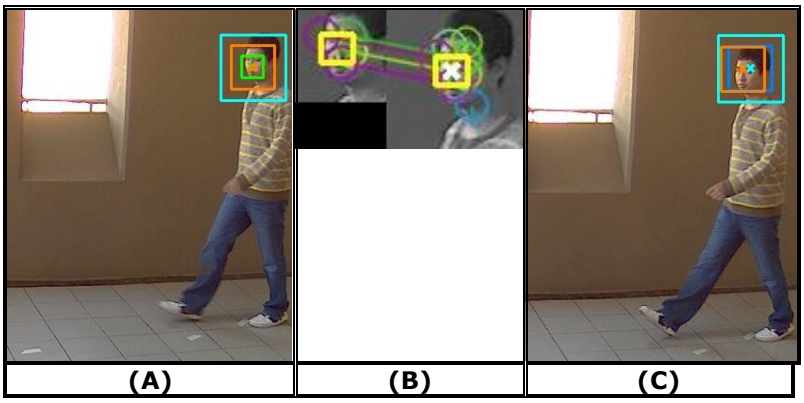
A operação do rastreador de faces SURF é dividida em duas etapas: (i) detecção de movimento; e (ii) *matching* de características SURF. Na

primeira etapa, a detecção de movimento é realizada pela subtração de *background*, em que o *background* é modelado em um domínio de padrões de *Bayer* (filtro monocromático RGB) pelo uso de uma mistura de gaussianas (MoG) (BISHOP, 2006) e o *foreground* é classificado em uma interpolação no domínio RGB.

Uma vez que uma face é detectada, o rastreador de faces determina uma região de busca tipicamente duas vezes maior que a região detectada da face com o propósito de encontrar a nova posição da face no quadro seguinte. Para realizar tal atividade, descritores de características SURF são extraídos da face detectada do quadro anterior e da região de busca.

Em seguida, tais descritores são comparados por meio de seus pontos de interesse com a utilização do algoritmo FANN – *Fast Approximate Nearest Neighbors* (MUJA e LOWE, 2009) que determina a nova posição da face no quadro seguinte, conforme ilustrado na Figura C.4.

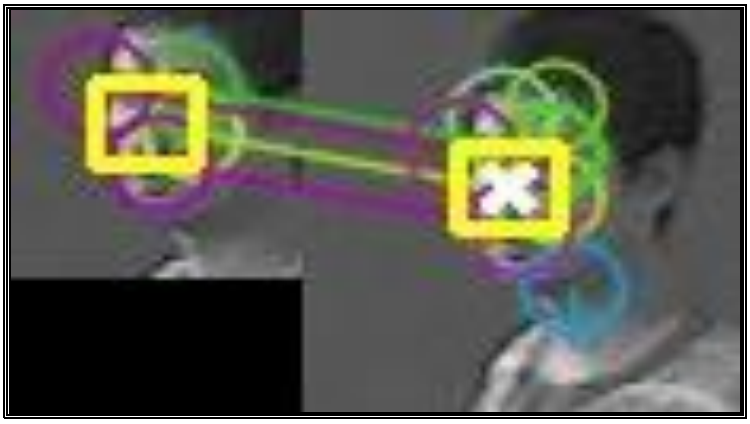
Figura C.4 – Exemplo de rastreamento de faces: (A) A região verde representa a face detectada, a região laranja representa uma vez o tamanho da região detectada e a região azul representa a região de busca (duas vezes o tamanho da região detectada); (B) Extração e comparação das características SURF; e (C) Determinação da nova posição central da face.



A determinação da face detectada, pelo algoritmo de rastreamento de faces, é feita a partir do centroide (x,y) da região que é delimitada pelo casamento dos pontos interesse. Este mecanismo pode ser visualizado pelo "X" de cor branca interno à região de cor amarela, conforme ilustrado na Figura C.5.

Neste exemplo, quatro correspondências foram destacadas (quatro linhas em verde escuro, verde claro e magenta). Para cada correspondência entre pontos de interesse, histogramas de cinza são extraídos da região do descritor.

Figura C.5 – Ilustração das regiões de comparação e das correspondências entre os pontos de interesse após a comparação dos respectivos descritores.



Assim, o algoritmo do rastreador de faces pode ser sumariado como:

Algoritmo C.1 – Rastreador de faces.

Entrada: Vídeo ou sequência de quadros contendo faces.
Inicialização: Localização central (x, y) da região de interesse e tamanho (w, h) da primeira face detectada.
Iteração: Para cada quadro consecutivo em que seja detectado movimento, os seguintes passos são realizados:

(1) Extração de descritores de características SURF da região da face detectada no quadro anterior;
(2) Extração de descritores de características SURF da região de busca do quadro atual (cujo tamanho é duas vezes o tamanho da região da face detectada);
(3) Comparação de descritores e pontos de interesse por meio do algoritmo FANN (*matching*);
(4) Determinação da localização (x,y) do centroide dos pontos de interesse correspondentes após o *matching*;
(5) Atualização da posição central (x, y) do rastreador de faces;
(6) Atualização do tamanho (w, h) da área rastreada com base na razão entre a região dos pontos de interesse do quadro anterior com a região dos pontos de interesse do quadro atual.

Saída: Localização central (x, y) e tamanho (w, h) da face de cada quadro (informação de entrada para geração de *face tracks*).

O algoritmo FANN utilizado no passo (3) do rastreador de faces é um método que aplica busca de prioridade em árvores *K-Means* hierárquicas (*Hierarchical K-Means Tree*), proposto por Muja e Loew (2009). A árvore *K-Means* hierárquica é construída dividindo-se os pontos de dados de cada nível em K regiões distintas por intermédio de um agrupamento *K-Means*. O mesmo procedimento é aplicado recursivamente para os pontos em cada região. A recursão é encerrada quando o número de pontos em uma região for menor do que o parâmetro *K*.

O algoritmo FANN explora a árvore *K-Means* hierárquica da forma *best-bin-first* (com base em árvores *kd*), ou seja, retorna o vizinho mais

próximo para uma grande fração de consultas e um vizinho muito próximo para os demais casos (HAR-PELED e KUMAR, 2010).

O algoritmo FANN realiza, inicialmente, um único percurso pela árvore, adicionando uma fila de prioridades a todos os ramos inexplorados em cada nó ao longo do caminho. Em seguida, extrai da fila de prioridades o ramo que possui o centro mais próximo ao ponto de consulta e reinicia a travessia da árvore a partir daquele ramo. Em cada passagem, o algoritmo continua adicionando à fila de prioridades os ramos inexplorados ao longo do caminho. O grau de aproximação é especificado da mesma forma como é definido para árvores *kd* randomizadas, parando a pesquisa logo após um número predeterminado de nós folha (pontos de dados) serem examinados (MUJA e LOEW, 2009).

Finalmente, um classificador *C* é utilizado para a verificação de um par de faces (A,B), e pode ser definido conforme a Equação (C.6).

$$C(A,B) = \begin{cases} Match & se\, S(A,B) \geq T \\ NoMatch & se\, S(A,B) < T \end{cases}, \quad \textbf{(C.6)}$$

em que *T* é limiar parametrizável, *Match* e *NoMatch* indicam se houve ou não uma correspondência entre as faces A e B, respectivamente. O parâmetro *T* pode ser empiricamente determinado de maneira a produzir o desempenho desejado, em termos de taxas de aceitação ou rejeição. Caso ocorra um casamento (*Match*), o rastreador de faces irá continuar o rastreamento da face corrente nos quadros seguintes. Caso contrário, um novo rastreador de faces é instanciado para rastrear a nova face encontrada.

C.5. Rastreador FRAG

O rastreador *Frag* baseia-se em fragmentos ou *patches*. Dado um objeto *O* no quadro corrente *I*, deseja-se localizar *O* na imagem *I*, em que, o objeto *O* é representado por um template *T*. Além disso, deseja-se encontrar a posição e a escala de uma região em *I* que seja a mais próxima de *T*, segundo um determinado critério. Estima-se, previamente, que a posição e a escala inicial do objeto sejam conhecidos e que a busca será realizada na

vizinhança desta estimativa.

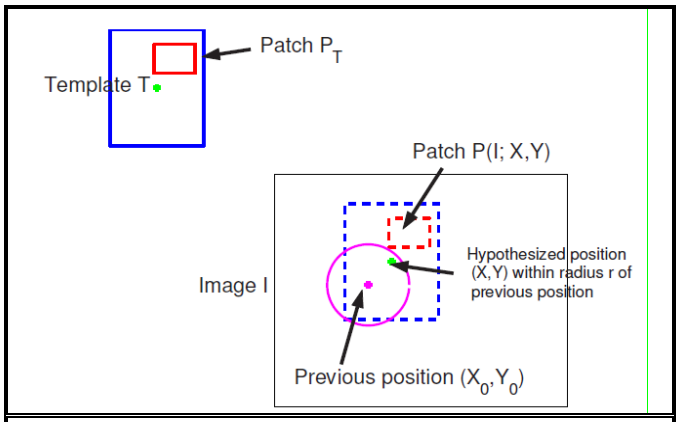
Portanto, o funcionamento de um rastreador com base em fragmentos (*Frag*) pode ser definido como descrito no Algortimo C.2.

Algoritmo C.2 – Rastreador de faces com base em fragmentos.

Entrada: Vídeo ou sequência de imagens contendo faces.
Inicialização: Localização central (x_0, y_0) estimada do objeto no quadro anterior e o raio *r* de busca.
Iteração: Para cada quadro consecutivo, os seguintes passos são realizados:
(1) Calcula-se $P_T = (dx, dy, h, w)$ de um fragmento retangular do *template*, cujo centro está deslocado (dx, dy) do centro do *template* com a **metade** da altura e largura sejam w e h, respectivamente;
(2) Seja (x, y) a posição hipotética do objeto no quadro corrente;
(2) Então, o fragmento P_T define um *patch* correspondente na imagem $P_I(x, y)$, cujo centro está no ponto $(x + dx, y + dy)$ com a metade da largura e altura, w e h.
Saída: Localização central (x, y) do objeto de cada quadro (informação básica para a geração de *face tracklets*).

Este mecanismo pode ser visualizado na Figura C.6.

Figura C.6 – *Template patch* P_t e o fragmento correspondente na imagem $P_I(x,y)$ para a posição hipotética (x,y).



Fonte: Adaptado de Adam, Rivlin e Shimshoni (2006).

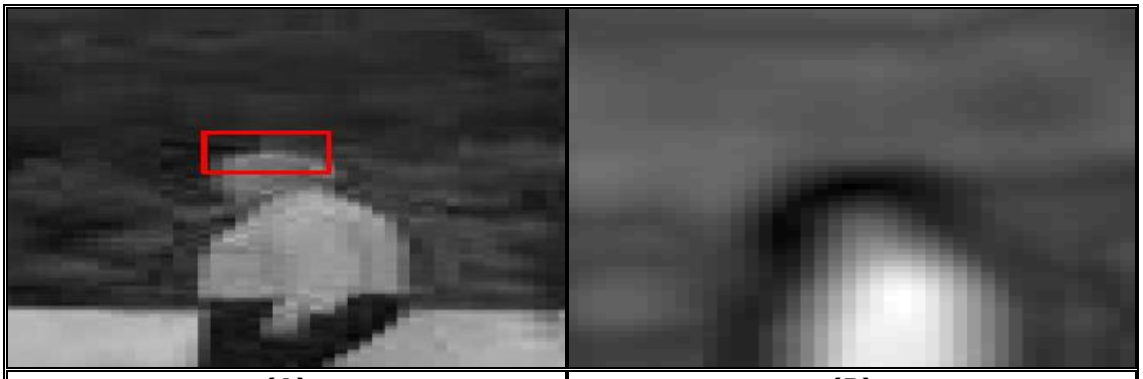
Dado o patch P_T e o correspondente $P_I(x, y)$, a similaridade entre eles é uma indicação da validade hipotética do objeto estar realmente localizado na posição (x, y). Adicionalmente, calcula-se uma medida de similaridade entre *patches* Q e P, denotada por $d(Q, P)$ e descrita por:

$$V_{Pt}(x, y) = d(P_I(x, y), P_t) \quad \text{(C.7)}$$

Quando (x, y) é considerado em um conjunto de hipóteses, obtém-se $V_{Pt}(\cdot, \cdot)$, que é o mapa de votação (V) correspondente ao *patch* do *template* P_T. Para o cálculo de d utiliza-se *Earth Mover's Distance* (**EMD**) entre dois histogramas. O valor de V é calculado para todos os locais ao redor do

centro do *patch* que são até 30 pixels acima ou abaixo e até 20 pixels para a esquerda ou para a direita, conforme ilustrado na Figura C.7.

Figura C.7 – (A) Exemplo de um *patch*. (B) Mapa de votação de similaridade EMD. Quanto mais escura a região, maior a possibilidade de indicar a posição estimada do objeto.



(A) **(B)**

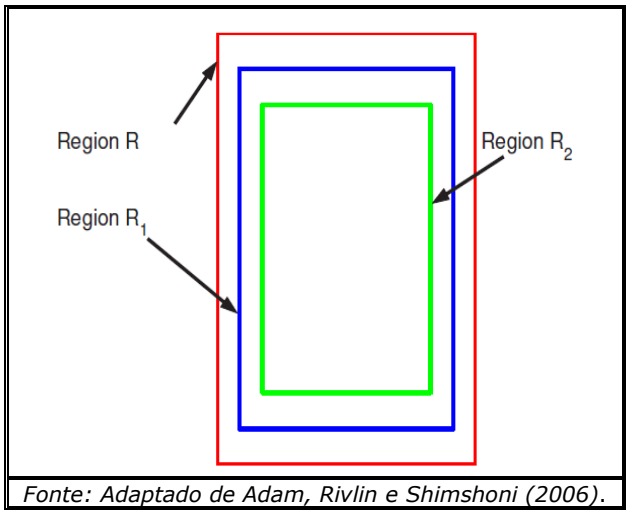
Fonte: Adaptado de Adam, Rivlin e Shimshoni (2006).

Para cada ponto ***(x, y)*** os valores de mapas de votação obtidos são ordenados {VP (x, y) | *patches* P} e o ***Q-ésimo*** menor é escolhido, conforme Equação (C.8):

$$C(x,y) = Q'th\,valor\,no\,conjunto\{VP(x,y) \mid patches\,P\} \quad \textbf{(C.8)}$$

Se ***Q*** for 25% do número de *patches*, então, pelo menos 1/4 do objeto será visível. Esta estratégia é benéfica para o tratamento de oclusões. Adicionalmente, existe a atribuição de menor peso para pixels que estão longe do centro do objeto, conforme ilustrado na Figura C.8. Este cálculo é realizado com uso de histogramas integrais.

Figura C.8 – Com base no uso de histogramas integrais, atribui-se menos peso a contribuições da parte exterior da região.

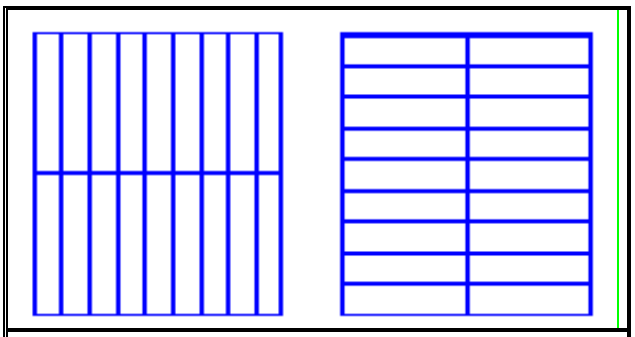


Fonte: Adaptado de Adam, Rivlin e Shimshoni (2006).

Em média, são utilizados 36 *patches* verticais de 1/2 da altura do *template* e de cerca de 1/10 da largura do *template* (vide Figura C.9), nos

quais são extraídos histogramas de cinza com 16 *bins*, um raio de busca de 7 pixels e ***Q*** = 25%. Em relação à escala, adota-se um aumento e diminuição de 10% do *template*. Por fim, são escolhidas a posição e escala que obtiverem o menor valor de ***C(x, y)***.

Figura C.9 – Patches utilizados pelo rastreador.



Fonte: Adaptado de Adam, Rivlin e Shimshoni (2006).

C.6. Rastreador TLD

O trabalho de Kalal, Mikolajczyk e Matas (2012), denominado *Tracking-Learning-Detection* – TLD, investiga o rastreamento a longo-termo de objetos em um vídeo. O método proposto é composto de três etapas, rastreamento, aprendizagem e detecção. Em particular, o rastreamento e a detecção são processos independentes que trocam informações usando um paradigma de aprendizagem P-N que estima os erros de detecção para otimizar o classificador.

A cada quadro, a aprendizagem P-N é realizada com os passos: (i) avaliação do detector no quadro corrente; (ii) estimativa dos erros de detecção utilizando especialista P-N; e (iii) atualização do detector pelas saídas rotuladas dos especialistas. Neste método, o especialista P-N é o elemento crucial que explora a estrutura temporal (*i.e.*, assume que o objeto se move ao longo de uma trajetória) e a estrutura espacial (i.e. assume que o objeto pode aparecer em um única localização) do vídeo, assim como, as respostas produzidas pelo rastreador quadro-a-quadro, para estimar erros de falsos negativos e falsos positivos do detector, respectivamente.

Experimentos realizados em três bases de imagens e em comparação com cinco outros métodos de rastreamento, demonstraram que o método

TLD é bastante promissor, obtendo 82% de *precision*, 81% de *recall* e 81% de *f-measure*.

C.7. Detector de Faces PICO

A ideia básica do detector de Faces PICO (*Pixel Intensity Comparisons Organized*) proposto por Markus et al. (2014) é varrer a imagem a partir de uma cascata de classificadores binários em diferentes tamanhos e escalas. Uma região da imagem é classificada como um objeto de interesse se a mesma passa com êxito por todos os membros da cascata.

Tal cascata, gradativamente, nível após nível, rejeita padrões de não faces, deixando para o último nível da cascata a confirmação da existência de uma face, com alta probabilidade. A ideia básica é sistematicamente percorrer a imagem de entrada com este esquema de classificação binária, alimentando-o com sub-regiões candidatas de diferentes tamanhos, escalas e posições. Cada classificador binário consiste de um conjunto de árvores de decisão com comparações de intensidade de pixel como testes binários em seus nós internos, permitindo processar rapidamente as regiões da imagem. Uma região da imagem é classificada como face se a mesma passa com êxito por todos os níveis da cascata.

O processo de aprendizagem do detector PICO consiste na construção de uma árvore de regressão gananciosa (*greedy*) e de um algoritmo de *boosting*. Experimentos realizados na base de referência FDDB (JAIN e LEARNED-MILLER, 2010) – com 5171 imagens de faces adquiridas em ambientes não controlados – obtiverem resultados superiores ao método de Viola e Jones (2001).

As vantagens do método PICO são: (i) não requer o cálculo de imagens integrais, imagem pirâmide ou outra estrutura de dados semelhante; e (ii) não requer pré-processamento, tais como, normalização de contraste, redimensionamento, correção de gama ou suavização gaussiana. Uma desvantagem é a necessidade de adaptação do método para detecção de faces com rotações no plano.

C.8. Detector de Faces CascadeCNN

O método desenvolvido na pesquisa de Li et al. (2015), denominado CascadeCNN - *Convolutional Neural Network Cascade for Face Detection* também faz uso de uma arquitetura em cascata, mas, ao invés de árvores de decisão, utiliza Redes Neurais Convolucionais (CNN), as quais possuem alta capacidade discriminativa. A cascata de CNN opera em múltiplas resoluções da imagem, rejeitando regiões de *background* nos estágios de baixa resolução e avalia cuidadosamente uma pequena quantidade de regiões de face candidatas no último estágio de alta resolução. A fim de melhorar a eficácia da localização final da face e reduzir o número de regiões candidatas nos últimos estágios, é realizada uma etapa de calibração depois de cada estágio de detecção da cascata. A saída de cada fase de calibração é utilizada para ajustar a posição da janela de detecção para a entrada para a etapa subsequente.

O experimento realizado por Li et al. (2015) na base de imagens FDDB (JAIN e LEARNED-MILLER, 2010) obteve uma taxa de detecção de 85,8% no escore discreto. Uma vantagem do método é que não requer pré-processamentos, tais como, normalização de contraste, redimensionamento, correção de gama ou suavização gaussiana. Uma desvantagem é a necessidade de calibração interna a cada estágio da cascata penalizando o tempo de execução do método.

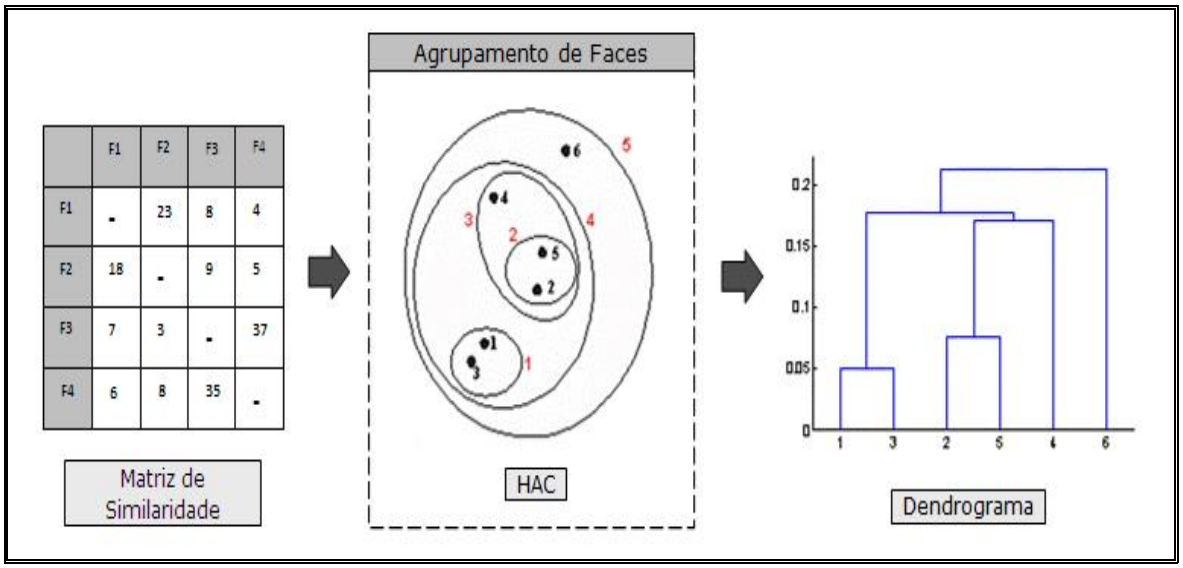
C.9. Agrupamento Hierárquico

Os métodos hierárquicos não requerem que seja definido um número *a priori* de agrupamentos. A partir da análise do dendrograma, diagrama que mostra a hierarquia e a relação dos agrupamentos em uma estrutura (ver Figura C.10), pode-se inferir o número de agrupamentos adequados.

Em um método hierárquico, inicia-se com cada padrão formando seu próprio agrupamento e, gradualmente, os grupos são unidos, até que um único agrupamento contendo todos os dados seja gerado. Logo no início do processo, os agrupamentos são pequenos e os elementos de cada grupo possuem um alto grau de similaridade. Ao final do processo, têm-se poucos

agrupamentos, cada um podendo conter muitos elementos e mais similares entre si (SILVA, 2005).

Figura C.10 – Exemplo de operação do método HAC.



Uma vez criada a matriz de similaridade pelo módulo de comparação e determinação de similaridade, o próximo passo é encontrar o menor valor da matriz. Esse valor representa a maior similaridade entre dois agrupamentos, os quais são agrupados, formando, assim, um novo agrupamento. Logo em seguida, a matriz de similaridade é atualizada, contendo agora um agrupamento a menos. Esse procedimento é repetido, até restar apenas um único agrupamento.

De acordo com Matteucci (2011), o procedimento geral de um método de agrupamento hierárquico pode ser formalizado como segue:

(1) Inicialmente, cada agrupamento contém um único padrão;

(2) Calcula-se/atualiza-se a matriz de similaridade;

(3) Forma-se um novo agrupamento pela união dos agrupamentos com maior grau de similaridade; e

(4) Os passos 2 e 3 são executados (N-1) vezes, até que todos os objetos estejam em um único agrupamento.

Diversos algoritmos hierárquicos foram propostos, dentre os quais destacam-se: (i) Agrupamento por Ligação Simples; (ii) Agrupamento por Ligação Completa; (iii) Agrupamento por Centroide; e (iv) Agrupamento de Ward (MATTEUCCI, 2011). No entanto, o algoritmo adotado neste trabalho

foi o Agrupamento por Média Aritmética Não-Ponderada (*Unweighted Pair-Group Method using Arithmetic Average* – UPGMA) que produziu melhores resultados em comparação aos anteriores nos experimentos realizados (Capítulo 4).

O agrupamento UPGMA é baseado na decisão de fundir dois agrupamentos em um cálculo que envolve as similaridades (ou proximidades) entre todos os objetos de ambos os agrupamentos analisados. Neste caso, a maior similaridade é encontrada pelo cálculo da média aritmética de todas as distâncias entre os objetos de um dos agrupamentos em relação, aos objetos do outro agrupamento (JAIN, 1991), conforme a Equação (C.9):

$$d(A,B) = \frac{1}{|A|.|B|} \sum_{x \in A} \sum_{y \in B} dist(x,y) \qquad \textbf{(C.9)}$$

Os agrupamentos que apresentarem o resultado de maior similaridade (ou proximidade) são agrupados.